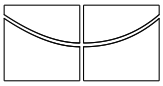

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Modelo de Resposta Gradual Para Testes Com Penalização Para Itens Dicotômicos.

por

Raquel Araújo de Almeida

Orientador: Prof. Dr. Antonio Eduardo Gomes

Novembro de 2015

Raquel Araújo de Almeida

Modelo de Resposta Gradual Para Testes Com Penalização Para Itens Dicotômicos.

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Novembro de 2015

Termo de Aprovação

Raquel Araújo de Almeida

Modelo de Resposta Gradual Para Testes Com Penalização Para Itens Dicotômicos.

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Data da defesa: 30 de Novembro de 2015

Orientador:

Prof. Dr. Antonio Eduardo Gomes
Departamento de Estatística, UnB

Comissão Examinadora:

Prof. Dr. Raul Yukihiro Matsushita
Departamento de Estatística, UnB

Prof. Dr. Paulo Henrique Portela de Carvalho
CEBRASPE e Departamento de Engenharia Elétrica, UnB

Brasília, Novembro de 2015

Ficha Catalográfica

ALMEIDA, RAQUEL ARAÚJO DE
Modelo de Resposta Gradual para testes com penalização para
itens dicotômicos., (UnB - IE, Mestre em Estatística, 2015).

Dissertação de Mestrado - Universidade de Brasília. Departamento de Estatística
- Instituto de Ciências Exatas.

1. Teoria da Resposta ao Item 2. Modelo de Resposta Gradual 3. Dados Faltantes
4. Vestibular 5. CEBRASPE 6. Avaliação

À minha mãe,

a quem dedico

todas as minhas vitórias e conquistas.

Agradecimentos

A Deus e à Nossa Senhora, por sempre estarem ao lado, me acompanhando e abençoando em todas horas, iluminando todos os meus caminhos e decisões.

À minha mãe, Nukácia, por tudo que me proporcionou e me proporciona até hoje. Pelo exemplo de mãe, mulher e profissional forte e batalhadora, da qual tenho muito orgulho. Por todo o apoio em forma de amor, carinho, conselhos, broncas, abraços, beijos e histórias para dormir. Sem a senhora eu não sou nada, sem a senhora eu não estaria aqui. Eu não tenho e não acho palavras que simbolizem e expressem o tamanho da minha gratidão. Muito obrigada!

À minha família, pela admiração e pelas palavras de incentivo sempre. Pelo amor que sempre recebo quando estou com vocês, sensação de felicidade que não tenho em nenhum outro lugar do mundo, de que sinto muitas saudades, mas que sempre vale muito apena retornar para sentir.

Ao meu amigo, companheiro, namorado, noivo ou marido (como ele prefere ser referido), Alan Cairo, pelo companheirismo inabalável. Por ser uma das pessoas que mais torce por mim e acredita no meu potencial. Por me acalmar de uma forma admirável e me fortalecer sempre que preciso com seu otimismo surpreendente. Obrigada por cuidar de mim, sempre com muito carinho e amor. Você é o melhor companheiro que a vida poderia ter me dado.

Às minhas amigas e colegas de batalha Kelly, Laura e Janaína desde a graduação na Universidade Federal do Ceará. Apesar de moramos longe umas das outras, estamos sempre torcendo, incentivando e ajudando-nos para que continuemos sempre traçando o nosso futuro com muita determinação e perseverança.

Aos meus amigos do trabalho, João Renato, Diego e Débora, pelo incentivo sempre, em especial à Carol, por ter divido comigo as angústias e vitórias dessa fase

de nossas vidas e por também sempre me ajudar nas disciplinas e no que fosse preciso; à Akina, por todos os ensinamentos que me passou, pelas palavras de incentivo e por toda a ajuda no que fosse preciso; à Patrícia, por toda sua disposição e generosidade em me ajudar e tirar minhas dúvidas e ao Roberto, que apesar de pouco tempo de convivência, esteve sempre disposto em ajudar, sendo sempre tão prestativo e generoso.

Ao professor João Maurício Araújo Mota, da Universidade Federal do Ceará, por ter me ajudado a ingressar no mestrado, me passando os seus ensinamentos sempre com muita paciência e incentivo. Admiro muito a sua forma de ensinar e o amor que tem por seus alunos.

Ao meu professor e orientador Antonio Eduardo Gomes, pela orientação, paciência e compartilhamento de seus conhecimentos, os quais são muitos, e pela agradável convivência.

Aos professores membros da banca examinadora, Paulo Henrique Portela e Raul Yukihiro, por aceitarem o convite para participar da defesa e pelos comentários e sugestões que ajudaram a melhorar a versão final da dissertação. Agradeço também a professora Cibele Queiroz por aceitar participar da banca de qualificação e pelas relevantes sugestões.

Ao IPEA, por possibilitar sempre a qualificação dos que lá estão e incentivar a área acadêmica.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

O presente trabalho tem como objetivo propor um método alternativo de pontuação para as questões do tipo A das provas do vestibular da UnB elaboradas pelo CEBRASPE. Esse novo método de pontuação consiste em aplicar o Modelo de Resposta Gradual (MRG), da Teoria da Resposta ao Item, ao considerar como uma alternativa de resposta a "não resposta" do aluno às questões do tipo A das provas do vestibular da UnB, nas quais os indivíduos ganham um ponto ao acertarem a questão, perdem um ponto ao errarem a questão e ganham zero pontos por não responderem a questão. Este método de pontuação foi proposto com o intuito de melhorar a avaliação dos indivíduos, visto que a teoria da resposta ao item apresenta inúmeras vantagens em relação ao método convencionalmente utilizado pelo CEBRASPE. Os resultados mostraram uma alta proximidade das notas estimadas pelo modelo de resposta gradual e das notas obtidas pelo método convencional. Dessa forma, pôde-se concluir que o método proposto é eficiente em avaliar os indivíduos com o aditivo das vantagens que a Teoria da Resposta ao Item oferece. Porém, tanto os resultados como algumas limitações evidenciaram oportunidades futuras de pesquisas sob a proposta apresentada, ainda não exaustivamente exploradas.

Palavras Chave: *: Teoria da Resposta ao Item, Modelo de Resposta Gradual, Dados Faltantes, Vestibular, CEBRASPE, Avaliação.*

Abstract

This paper aims to propose an alternative scoring method for questions of type A of the UnB vestibular entrance exam prepared by CEBRASPE. This new correction method applies the Graded Response Model (GRM) from Item Response Theory. We consider "no answer" as an additional response class to the regular possibilities "right" or "wrong". in the UnB vestibular entrance exam, individuals score 1 for each question with right answer, -1 for each question with wrong answer, and 0 for no answer. This scoring method has been proposed in order to improve the evaluation of individuals' proficiencies, as the item response theory has many advantages compared to the conventional method used by CEBRASPE. The results showed a high correlation between the scores obtained with the use of the graded response model and the scores given by the conventional method. We concluded that the proposed method is effective in evaluating individuals' proficiencies with the advantages of the Item Response Theory. However, the results show opportunities for future research in the studied problem.

Key words: *Item Response Theory, Graded Response model, Missing data, Entrance exam, CEBRASPE, Evaluation.*

Capítulo 1

Introdução

Na maioria das pesquisas em que se trabalha com banco de dados é comum se deparar com dados faltantes. Esse tipo de problema tem se tornado cada vez mais frequente em diferentes áreas de pesquisas científicas como ciências sociais, educação, saúde, entre outras. A ocorrência de dados faltantes é uma limitação bastante delicada, tornando um desafio o uso do bancos de dados incompletos. Isso se dá, entre outros motivos, pelo fato de que a maioria das técnicas estatísticas são desenvolvidas para serem utilizadas em matrizes de dados completas. Assim, diferentes técnicas de tratamentos de dados faltantes vêm sendo desenvolvidas ao longo das últimas décadas.

Nesta dissertação, estudaremos o problema de dados faltantes em exames em que há penalização por resposta incorreta, como por exemplo, questões do tipo A das provas do vestibular da UnB realizado pelo CEBRASPE (Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos). Nas questões do tipo A deste tipo de exame, o respondente ganha 1 ponto quando acerta uma questão, perde 1 ponto quando erra e ganha zero pontos quando não responde, gerando assim um dado faltante. Neste trabalho propomos o cálculo da nota do respondente utilizando o modelo de resposta gradual da Teoria de Resposta ao Item (TRI), considerando três categorias: "Errar", "Não responder" e "Acertar". Este modelo considera que as categorias de resposta de uma questão podem ser ordenadas entre si, de tal forma que a categoria mais baixa contribua menos para o escore do respondente e a categoria mais alta contribua mais.

Sendo assim, iremos elaborar um sistema de pontuação, com pontuação di-

ferente para cada questão, mas considerando a categoria adicional de "não resposta" como uma categoria intermediária, visto que ela não é tão desvantajosa para o indivíduo como a categoria "Errar", mas também não é tão vantajosa para o indivíduo como a categoria "Acertar".

Neste sentido, o presente trabalho propõe comparar a nota obtida pelo modelo de resposta gradual com a nota calculada convencionalmente, de forma padronizada, o qual consiste em somar as pontuações dos respondentes para todas as questões do teste, subtraindo pela média de todas as notas e dividindo pelo desvio padrão, resultando assim, numa nota final, analisando, por exemplo, se há correlação entre as notas e quais as vantagens e desvantagens de se modelar as respostas destes tipos de testes através do modelo de resposta gradual da TRI.

Para implementarmos a análise proposta, iremos simular um banco de dados de respostas de indivíduos com dados faltantes a partir da definição de algumas regras. Além disso, iremos também, aplicar a metodologia sugerida nesta dissertação a um banco de dados real do vestibular da UnB.

Este trabalho está organizado da seguinte forma: Nos capítulos 2 e 3 estão apresentados a parte teórica utilizada neste trabalho, em que o capítulo 2 corresponde à abordagem da Teoria de Resposta ao Item e o capítulo 3 corresponde ao tema Dados Faltantes. No capítulo 4, é apresentada a metodologia deste trabalho, em que detalha como a análise foi realizada, tanto para os dados simulados, como para os dados reais. Em seguida, nos capítulos 5 e 6, estão expostos as análises e os resultados com relação aos dados simulados e aos dados reais, respectivamente. Por fim, no capítulo 7, é apresentada a conclusão deste trabalho, juntamente com as sugestões de trabalhos futuros.

Capítulo 2

Teoria da Resposta ao Item: alguns conceitos básicos

2.1 Contextualização

Em muitos estudos sociológicos, psicológicos ou educacionais a variável de interesse é de compreensão intuitiva para todos. Entretanto, na maioria dos casos, essa variável não é observável diretamente. Na área de psicometria, esses tipos de variáveis são chamadas de variáveis não observáveis, habilidades ou traços latentes.

Embora esse tipo de variável possa ser facilmente descrita através de suas características e conceitos, como inteligência, ansiedade, insegurança, entre outras, elas não podem ser medidas de forma direta, ao contrário de variáveis como peso ou altura de um indivíduo, mesmo que todas elas sejam características implícitas de cada ser humano.

A partir disso, quando se tem o interesse em medir um traço latente, torna-se necessário criar uma escala de medida, tal que essa variável assumirá valores contidos nessa escala. Entretanto, definir essa escala de medida, o intervalo desta escala e sua interpretação, em relação ao traço latente medido, são tarefas bastante difíceis por vários motivos técnicos.

A preocupação em medir traços sociológicos, educacionais e principalmente psicológicos é bastante antiga. Sendo assim, muitos estudos e propostas de métodos foram desenvolvidos no sentido de alcançar este objetivo. Uma das primeiras técnicas

utilizadas nessa área foi a Teoria Clássica dos Testes (TCT), também chamada de Teoria Clássica das Medidas (TCM), sendo esta, bastante desenvolvida já na década de 1950, principalmente com os trabalhos de Guilford (1954) e Gulliksen (1950).

Utilizando essa técnica, em um processo de avaliação e seleção de indivíduos, na área educacional, por exemplo, os respondentes são analisados tradicionalmente a partir dos resultados obtidos da aplicação de provas ou testes, expressos apenas por seus escores brutos ou padronizados. Dessa forma, a classificação de um indivíduo será melhor tanto quanto maior for sua nota na prova ou no teste. Esse procedimento caracteriza-se pelo fato de que as análises e interpretações estão sempre associadas ao escore total e não a um item ou questão em particular, sendo esta a principal característica da Teoria Clássica das Medidas.

No entanto, a TCM apresenta várias limitações. O fato de os resultados encontrados dependerem do particular conjunto de itens (questões) que compõem o instrumento de medida faz com que as análises e interpretações estejam sempre associadas à prova como um todo, tornando-se inviável a comparação entre indivíduos que não foram submetidos às mesmas provas.

Com o intuito de corrigir essas limitações da TCM, surgiu a técnica Teoria da Resposta ao Item (TRI), que propõe modelos de medição para traços latentes com algumas vantagens em relação à técnica tradicionalmente utilizada. A TRI é uma teoria desenvolvida não com o intuito de substituir a TCM, mas sim preencher lacunas, e assim adicionar a essa importante técnica novas possibilidades e áreas de aplicações.

De acordo com Araújo et al. (2009), existem várias vantagens da TRI em relação à Teoria Clássica. Uma das grandes vantagens é que ela possibilita fazer comparações entre traço latente de indivíduos de populações diferentes quando são submetidos a testes ou questionários que tenham alguns itens comuns e permite, ainda, a comparação de indivíduos da mesma população submetidos a testes totalmente diferentes. Isto é possível porque a TRI tem como elementos centrais os itens e não o teste ou questionário como um todo.

Outra vantagem da TRI é o fato de ela possibilitar uma melhor análise de cada item que forma o instrumento de medida, pois leva em consideração suas características específicas de construção de escalas. Além disso, os itens e os indivíduos

estão na mesma escala, com isso o nível de uma característica que um indivíduo possui pode ser comparado ao nível da característica exigida pelo item, facilitando assim a interpretação da escala gerada e permitindo também conhecer quais itens estão produzindo informação ao longo da escala. Outra vantagem apresentada pelos autores é que a TRI permite um tratamento para um conjunto de dados faltantes, utilizando para isso somente os dados respondidos, o que não pode acontecer na Teoria Clássica de Medidas. Pode-se citar também que outro benefício da TRI é o princípio da invariância, isto é, os parâmetros dos itens não dependem do traço latente do respondente, assim como os parâmetros dos indivíduos não dependem dos itens apresentados.

2.2 Breve Histórico

Segundo Andrade et al. (2000), os primeiros modelos da resposta ao item surgiram na década de 1950, e eram modelos em que se considerava que uma única habilidade, de um único grupo, estava sendo medida por um teste em que os itens eram medidos de maneira dicotômica.

Os primeiros modelos que surgiram consideravam apenas um traço latente a ser medido, denominados modelos unidimensionais. Inicialmente, surgiu o modelo unidimensional de dois parâmetros (dificuldade e discriminação) baseado na distribuição ogiva normal (normal acumulada), desenvolvido por Frederic Lord, principal estatístico do *Educational Testing Service*-ETS, nos Estados Unidos. Após algumas aplicações deste modelo, Lord (1952) incorporou ao modelo mais um parâmetro com o intuito de tratar o problema do acerto casual, surgindo assim o modelo unidimensional de três parâmetros.

Em 1968, Birnbaum desenvolveu uma proposta de substituição da função ogiva normal pela função logística, sendo esta utilizada até os dias de hoje. A função logística tem a vantagem de não envolver integração, além de ser uma função explícita dos parâmetros dos itens e do traço latente medido, sendo assim, matematicamente mais conveniente.

Paralelamente ao trabalho de Lord, Georg Rasch começou a desenvolver modelos de medida de traços latentes desde a década de 1940 e criou o modelo unidimensional de 1 parâmetro, conhecido também como Modelo de Rasch. Esse modelo

inicialmente foi expresso pela função ogiva normal e também mais tarde descrito através de uma função logística por Wright (1968).

No ano seguinte, Samejima (1969) apresentou o modelo de resposta gradual, a partir da necessidade de se introduzir nos testes respostas que não fossem classificadas exclusivamente como dicotômicas. Além disso, o autor teve como objetivo obter mais informações das respostas dos indivíduos do que simplesmente se eles deram respostas certas ou erradas às questões. Outros modelos para respostas politômicas, nominais ou graduais, foram propostos nos anos posteriores, tais como: o Modelo de Resposta Nominal de Bock (1972), o Modelo de Crédito Parcial proposto por Masters (1982), entre outros.

Mais tarde, modelos que permitem a análise e comparação de rendimentos de duas ou mais populações submetidas a testes diferentes, mas com itens em comum, começaram a ser desenvolvidos. Bock e Zimowski (1997) introduziram os modelos logísticos de 1, 2 e 3 parâmetros com mais de duas populações.

A TRI foi utilizada inicialmente no Brasil em 1995, na análise dos resultados do Sistema Nacional de Ensino Básico (SAEB) e posteriormente foi utilizada também pelo Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP). Com a complexidade matemática dos cálculos que são necessários para a aplicação dessa teoria, o desenvolvimento de pacotes computacionais veio viabilizar a utilização da TRI em grande escala, começando sua aplicação sistemática por volta de 1980, nos Estados Unidos, em estudos de avaliação educacional.

Por consequência, essa técnica tem despertado interesse de aplicação em diferentes áreas, tais como: na medicina, quando Das e Hammer (2005, apud BOSI, 2010) desenvolveram um estudo na Índia, com o intuito de avaliar a diferença entre os cuidados de saúde oferecidos nas regiões pobres e ricas; na psicologia, quando Andriola (1998, apud BOSI, 2010) criou um banco de itens com o objetivo de avaliar a capacidade cognitiva em alunos do ensino médio; no campo do marketing, na avaliação de usabilidade em sites de e-commerce, desenvolvido por Tezzaet et al. (2009, apud BOSI, 2010); na área social, quando Costa et al. (2009, apud BOSI, 2010) analisaram os efeitos do Programa Jovens Baianos de Formação de Agentes e Desenvolvimento Comunitário (ADCs), que teve como objetivo proporcionar ao jovem oportunidades de acesso e permanência na escola, inclusão produtiva e de empreendimentos de ações

comunitárias; na área de produção e índices sócio-econômicos por Soares (2005); na gestão pela qualidade total por Alexandre et al. (2002), entre outros.

2.3 Modelos da Teoria da Resposta ao Item

Segundo Demars (2010), modelos da Teoria da Resposta ao Item mostram a relação entre a habilidade ou característica medida pelo instrumento e uma resposta a um determinado item. As respostas dos itens podem ser classificadas como dicotômicas (duas categorias), tais como certo ou errado, sim ou não, concordar ou discordar; ou podem ser classificadas como politômicas (acima de duas categorias), tais como a classificação de um serviço ou opções de respostas na escala de Likert de uma determinada pesquisa.

Andrade et al. (2000) definem a TRI como o conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros dos itens e da habilidade ou habilidades do respondente, dependendo do modelo em questão. O conjunto de modelos foi desenvolvido de forma que, quanto maior a habilidade do indivíduo, maior a probabilidade de ele dar uma resposta correta ao item em questão.

Tais modelos dependem essencialmente de três atributos: natureza do item, sendo esta classificada como dicotômica ou não dicotômica, como apresentado anteriormente; número de populações envolvidas no modelo, discriminado em apenas uma ou mais de uma população; e, a quantidade de traços latentes que estão sendo medidos pelo modelo, classificados como unidimensional (apenas um traço latente medido) ou multidimensional (mais de um traço latente). Neste trabalho, serão abordados modelos unidimensionais, dicotômicos e não dicotômicos, com somente uma população envolvida.

2.3.1 Modelos para Itens Dicotômicos

Normalmente, os modelos mais utilizados da TRI são os modelos logísticos para itens dicotômicos. Basicamente, há três tipos de modelos nesse contexto, os quais se diferenciam pela quantidade de parâmetros que é utilizada para descrever o

item.

O modelo logístico unidimensional de 1 parâmetro (ML1), também conhecido como modelo de Rasch, é dado pela Equação 2.1 e possui apenas 1 parâmetro, como o próprio nome indica, o qual é chamado de parâmetro de dificuldade do item (b_i). O modelo logístico unidimensional de 2 parâmetros (ML2) é representado através da Equação 2.2. Além do parâmetro de dificuldade do item (b_i), o modelo é composto também pelo parâmetro de discriminação do item (a_i). Por último, o modelo logístico unidimensional de 3 parâmetros (ML3) é dado pela Equação 2.3, o qual é constituído também, além dos dois parâmetros do modelo anterior, pelo parâmetro de acerto casual (c_i).

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-D(\theta_j - b_i)}}, \quad \text{com} \quad i = 1, 2, \cdots, I, \quad \text{e} \quad j = 1, 2, \cdots, n; \quad (2.1)$$

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \quad \text{com} \quad i = 1, 2, \cdots, I, \quad \text{e} \quad j = 1, 2, \cdots, n; \quad (2.2)$$

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \quad \text{com} \quad i = 1, 2, \cdots, I, \quad \text{e} \quad j = 1, 2, \cdots, n; \quad (2.3)$$

onde:

- U_{ij}, é uma variável dicotômica, que assume o valor 1(um) quando o j-ésimo indivíduo responde corretamente o i-ésimo item, ou assume 0 (zero) quando o j-ésimo indivíduo não responde corretamente o i-ésimo item;
- θ_j, representa a habilidade do j-ésimo indivíduo;
- $P(U_{ij} = 1|\theta_j)$, é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente o item i. Essa probabilidade também é chamada de Função de Resposta do Item (FRI);
- b_i é o parâmetro de dificuldade, ou de posição, do item i, medido na mesma escala que a habilidade;

- a_i é o parâmetro de discriminação, ou de inclinação, do item i, com valor proporcional à inclinação da Curva Característica do Item (CCI) no ponto b_i;

- c_i é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item i (muitas vezes referido como a probabilidade de acerto casual);

- D é um fator de escala, constante igual a 1. Quando se deseja assemelhar os resultados desse modelo com os resultados obtidos por intermédio da função ogiva normal, é utilizado D=1,7.

Observa-se que os primeiros dois modelos podem ser obtidos facilmente a partir do modelo logístico de três parâmetros. Se, por acaso, não houver possibilidade de acerto ao acaso, considera-se o valor do parâmetro c igual a zero, e obtém-se assim o ML2. E se tivermos ainda, além da não possibilidade de resposta ao acaso, que os itens tenham o mesmo poder de discriminação, obtém-se o ML1.

2.3.1.1 Representação Gráfica: Curva Característica do Item

Como já citado anteriormente, a TRI é utilizada para analisar um conjunto de dados oriundo de respostas a itens de ferramentas avaliativas de desempenho ou questionários, propondo formas de representar a probabilidade de um indivíduo dar uma determinada resposta a um item, levando em conta suas proficiências ou habilidades e algumas características do item. Assim, a probabilidade $P(U_{ij} = 1|\theta_j)$ pode ser vista como a proporção de respostas corretas a um determinado item i dentre todos os indivíduos com proficiência θ_j. Graficamente, a relação existente entre $P(U_{ij} = 1|\theta_j)$ e os parâmetros do modelo é representada pela Curva Característica do Item (CCI), ilustrada na Figura 2.1.

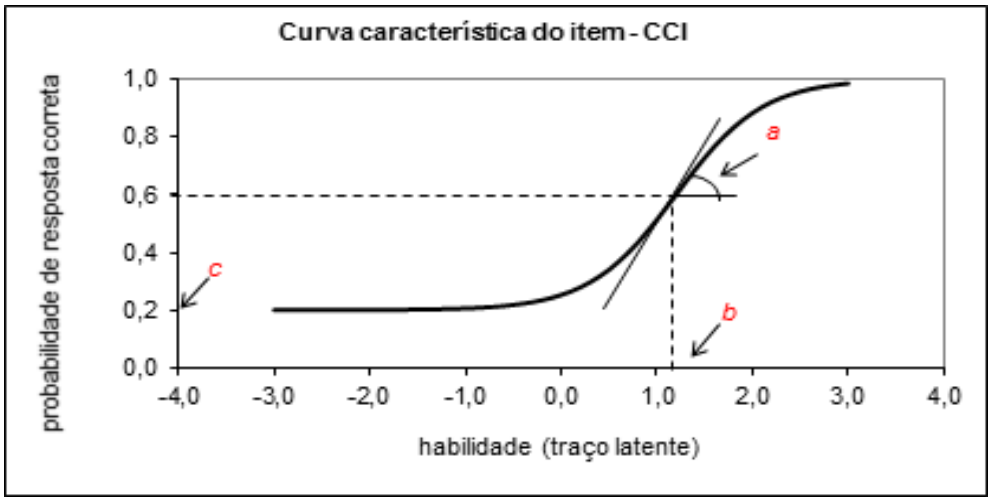

Figura 2.1: Exemplo de uma Curva Característica do Item (CCI) com $a_i = 1.3$, $b_i = 1.2$ e $c_i = 0.2$.
Fonte: Andrade et al. (2000).

Através da figura anterior, é possível observar que a curva CCI tem uma forma de "S", representando assim a relação não linear a qual o modelo se baseia, sendo esta definida como: *indivíduos com maior habilidade possuem maior probabilidade de acertar o item.* Em outras palavras, quanto mais alta for a habilidade ou proficiência do respondente em relação ao que está sendo medido no instrumento, maior será a probabilidade de resposta correta ao item. Assim, cada item possui uma CCI própria, com suas referidas características de dificuldade, discriminação e probabilidade de acerto casual, representadas pelos valores de seus respectivos parâmetros (b_i, a_i e c_i).

Parâmetros dos Itens

No que diz respeito à interpretação de cada um dos parâmetros dos itens, esta pode ser feita com o auxílio de figuras, como será mostrado a seguir.

O parâmetro de dificuldade b, como o próprio nome sugere, expressa o quão difícil é o item. Assim, quanto maior o valor de b, mais difícil o item será. Este parâmetro é medido na mesma escala da habilidade e pode ser visto como a habilidade necessária para que a probabilidade de acerto casual seja igual a $(1 + c)/2$. Segundo Baker (2001), um item fácil funciona bem entre examinados de baixa habilidade e um item difícil funciona bem entre examinados de alta habilidade. Assim, o parâmetro

de dificuldade é um índice de localização ou de posição.

De acordo com Demars (2010), o valor de b é igual ao valor da habilidade θ no ponto em que a inclinação da função é máxima. A Figura 2.2 apresenta dois itens com diferentes valores deste parâmetro. Ao se compararem os itens, a habilidade requerida para uma probabilidade de resposta correta de $0,60$, por exemplo, é igual a -1 para o item 2 e igual a 1 para o item 1. Isto é, o item 1 é mais difícil que o item 2. Assim, pode-se perceber que itens mais difíceis, ou seja, com maiores valores de b, exigem uma habilidade maior dos respondentes em relação ao que está sendo medido.

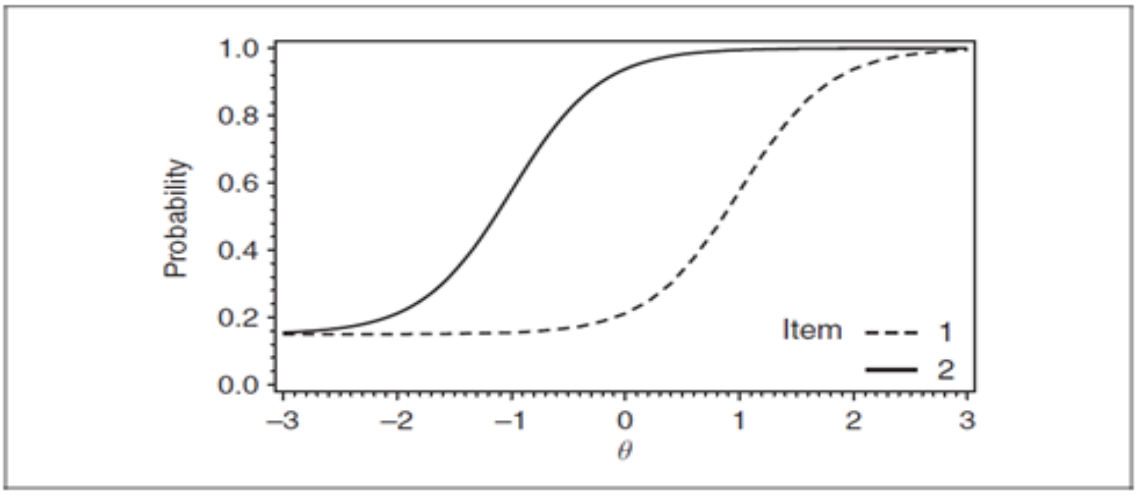

Figura 2.2: Curva Característica de dois itens com diferentes valores de b.
Fonte: Demars (2010).

O parâmetro de discriminação a descreve o quão bem um item pode diferenciar os examinados de acordo com suas respectivas habilidades em relação à proficiência medida. Este parâmetro é proporcional à derivada da tangente da curva no ponto de inflexão. Dessa forma, itens com a negativo não são esperados nesses modelos, visto que esses valores indicariam que a probabilidade de um indivíduo responder o item corretamente diminuiria com o aumento da habilidade. Segundo Baker (2001), quanto mais íngreme for a curva, melhor o item discrimina. Por outro lado, quanto mais plana a curva, menos o item é capaz de discriminar, uma vez que a probabilidade de resposta correta em baixos de níveis de habilidade é quase a mesma para altos níveis de habilidade.

Quanto maior a discriminação, ou seja, quanto maior o valor deste parâmetro, melhor o item discrimina os indivíduos avaliados com diferentes níveis de habilidades. Assim, se é desejada uma elevada discriminação. Andrade et al. (2000, p.11) afirmam que

> Baixos valores de a indicam que o item tem pouco poder de discriminação (alunos com habilidades bastante diferentes têm aproximadamente a mesma probabilidade de responder corretamente ao item) e valores muito altos indicam itens com curvas características muito "íngremes", que discriminam os alunos basicamente em dois grupos: os que possuem habilidades abaixo do valor do parâmetro b e os que possuem habilidades acima do valor do parâmetro b.

A Figura 2.3 mostra dois itens com diferentes valores do parâmetro a. Observa-se que o item 1 tem uma discriminação maior do que o item 2 por ter uma inclinação mais acentuada, ou seja, por ter uma CCI mais íngreme. Por este motivo, esse parâmetro também é chamado de parâmetro de inclinação do item. Numericamente, esta afirmação se comprova ao se observar que a diferença entre as probabilidades de resposta correta com habilidade 0 e 1, por exemplo, é maior no item 1 (aproximadamente $0,95 - 0,55 = 0,4$) do que no item 2 (aproximadamente $0,75 - 0,55 = 0,2$).

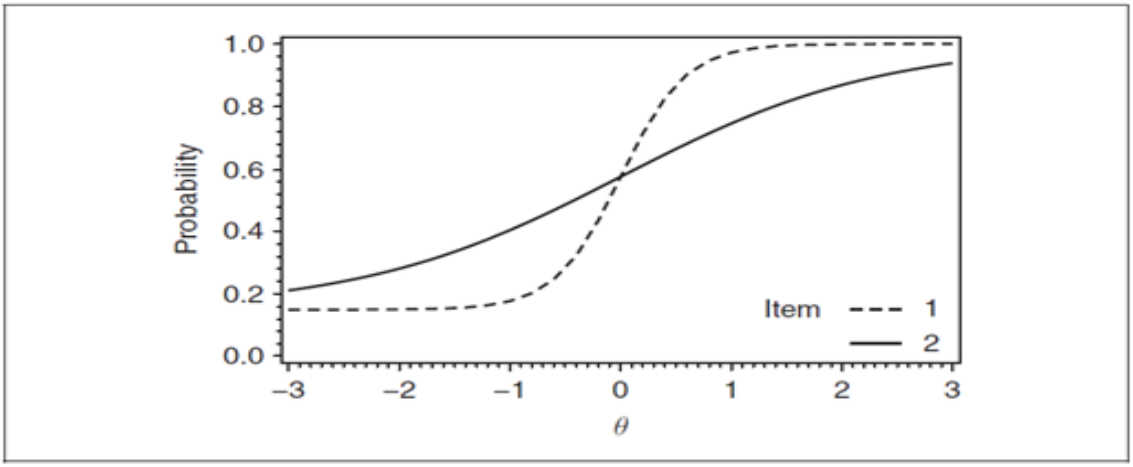

Figura 2.3: Curva Característica de dois itens com diferentes valores de a.
Fonte: Demars (2010).

O parâmetro c, conhecido como parâmetro de acerto casual, representa a probabilidade de um aluno com baixa habilidade responder corretamente o item. Por conta disso, este parâmetro é chamado também muitas vezes de parâmetro de "adivinhação". Este parâmetro não depende de escala, pois trata-se de uma probabilidade, assumindo, assim, sempre valores entre 0 e 1. De acordo com Baker (2001), é importante notar que o parâmetro c não varia como uma função do nível da habilidade. Assim, tanto examinados de alta como de baixa habilidade têm a mesma probabilidade de adivinhar a resposta correta do item.

De acordo com Santos (2009, p. 4),

> Se um item de múltipla escolha é construído de tal forma que as alternativas incorretas (distratores) funcionem muito bem, ou seja, se os distratores cumprem o seu papel de trazer informação ao avaliador a respeito da manifestação do raciocínio do aluno quando busca a solução para a tarefa imposta pelo item, mas sem chamar mais atenção do que a resposta correta, provavelmente o parâmetro c estará em torno do inverso do número de alternativas. Todavia, na prática, observam-se valores desde próximos de zero até próximos a 0,5, raramente ultrapassando esse valor.

Quando não é permitido "chutar", o parâmetro c assume valor 0 e o parâmetro b representa o ponto na escala da habilidade onde a probabilidade de acertar o item é 0,5. Segundo Demars (2010), a assíntota a esquerda ou assíntota inferior é utilizada para descrever o valor que a função assume quando o valor de θ se aproxima de menos infinito. Assim, o parâmetro c pode ser visto também como a assíntota inferior da curva característica do item. A Figura 2.4 mostra a curva característica de dois itens com diferentes valores do parâmetro c. Quanto menor é a assíntota, menor o valor de c, sendo assim, no item 1, um individuo com baixa habilidade tem menor probabilidade de acertar o item do que no item 2.

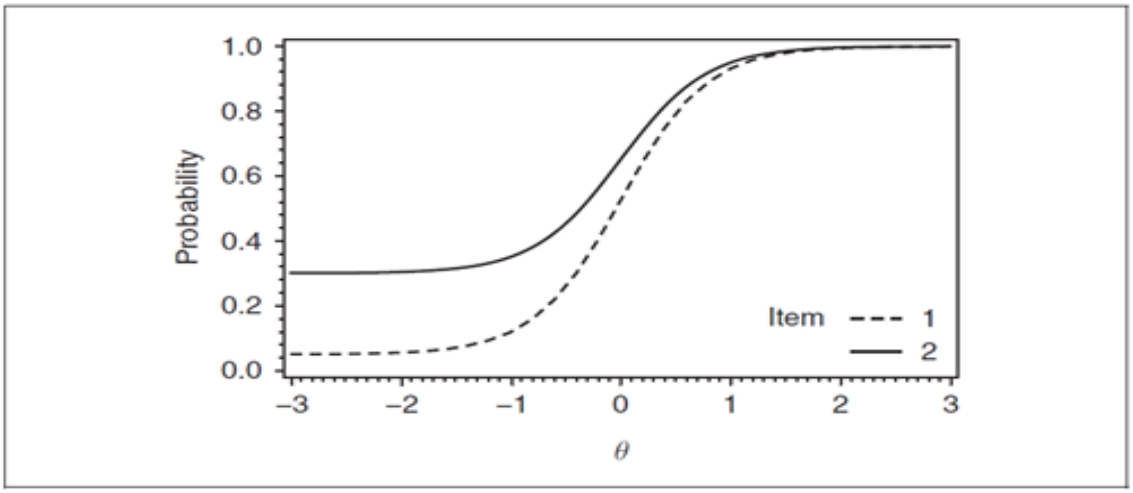

Figura 2.4: Curva Característica de dois itens com diferentes valores de c.
Fonte: Demars (2010).

Escala da Habilidade

O parâmetro de habilidade θ pode assumir qualquer valor entre $-\infty$ a $+\infty$. Logo, é necessário estabelecer uma origem, de modo a representar o valor médio (μ) das habilidades, e uma unidade de medida, representando o desvio-padrão (σ) das habilidades dos indivíduos da população em estudo. Normalmente, segundo Andrade et al. (2000), utiliza-se a escala com média igual a 0 e desvio padrão igual a 1, sendo esta representada por escala $(0, 1)$. Nessa escala, portanto, os valores do parâmetro b variam, tipicamente, entre -2 e +2 e os valores do parâmetro a variam entre 0 e +2, em que vale ressaltar que os valores mais adequados de a seriam valores maiores que 1.

Apesar de ser comum a utilização da escala (0,1), na prática, não há diferença entre escolher estes valores ou quaisquer outros para a escala da habilidade. O que realmente importa são as relações de ordem existentes entre seus pontos. Isso significa dizer que, independentemente da escala utilizada para medir a habilidade do indivíduo, a probabilidade de ele responder corretamente a um certo item é sempre a mesma, ou seja, a habilidade de um respondente é invariante à escala de medida.

Na prática, as habilidades e os parâmetros dos itens são estimados a partir das respostas de um grupo de indivíduos submetidos a esses itens. Entretanto, uma vez estabelecida a escala de medida do traço latente, os valores dos parâmetros dos

itens não mudam. Em outras palavras, os valores de a e b são invariantes a diferentes grupos de respondentes, desde que os indivíduos destes grupos tenham suas habilidades medidas na mesma escala definida. Sendo assim, não há sentido algum em tentar analisar os itens a partir dos valores de seus parâmetros a e b, sem antes conhecer a escala na qual eles foram definidos.

2.3.1.2 Função de Informação do Item e Função de Informação do Teste

Segundo van der Linden e Hambleton (1997), uma das outras contribuições de Birnbaum à teoria psicométrica foi a introdução da medida de Fisher para descrever a estrutura de informação de um teste. Mais conhecida como Função de Informação do Item (FII), essa medida é bastante utilizada em conjunto com a CCI. Ela permite verificar o quanto um item, ou um teste, contém de informação para a medida de habilidade. A FII é dada por:

$$I_i(\theta) = \frac{[\frac{d}{d\theta}P_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \tag{2.4}$$

em que

$I_i(\theta)$ é a "informação" fornecida pelo item i no nível de habilidade θ;
$P_i(\theta)$= $P(X_{ij} = 1|\theta)$ é a probabilidade de acerto ao item dado θ;
$Q_i(\theta) = 1 - P_i(\theta)$;
$\frac{\partial P_i(\theta)}{\partial \theta}$ é a derivada de $P_i(\theta)$

Em particular, no modelo logístico de três parâmetros a equação pode ser escrita como:

$$I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2. \tag{2.5}$$

Esta equação mostra a importância que têm os três parâmetros dos itens sobre o montante de informação do item. Na figura a seguir (Figura 2.5), estão representadas a Curva Característica do Item e a Curva de Informação do Item de um item hipotético.

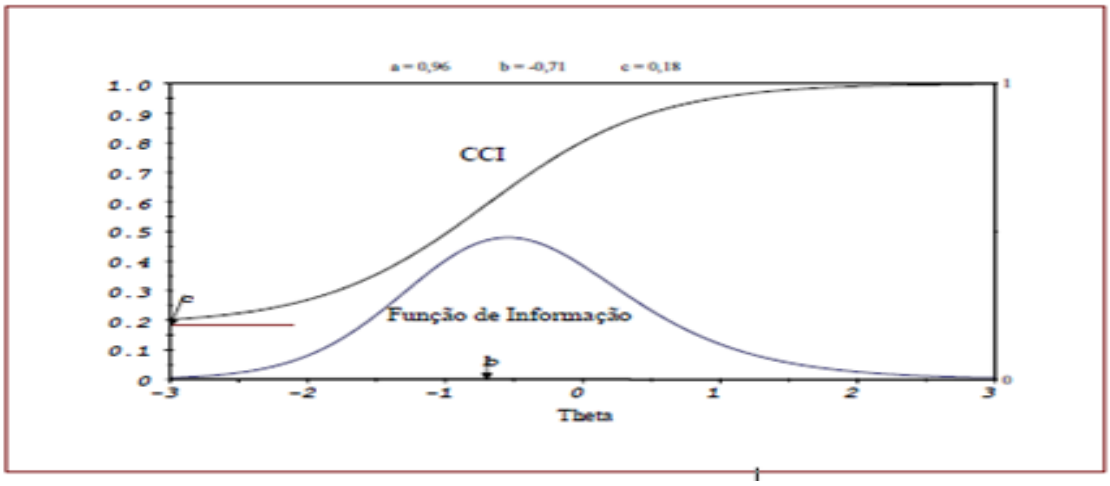

Figura 2.5: Curva Característica do item e sua respectiva Curva de Informação do item.
Fonte: Andriola(2009).

Pode-se observar que o item proporciona maior quantidade de informação nos valores de θ dentro do intervalo de -1,5 a 0,5. De acordo com Andrade et al. (2000), a informação é maior segundo três condições:

a) Quando b_i se aproxima de θ;

b) Quanto maior for o valor de a_i;

c) Quanto mais c_i se aproxima de 0;

Segundo ainda os mesmo autores, percebe-se que cada item está associado a um intervalo na escala de habilidade no qual o item tem maior poder de discriminação. Este intervalo é definido em torno do valor do parâmetro b e está mostrado pela curva de informação do item. Logo, a discriminação entre bons respondentes é feita a partir de itens considerados difíceis e não itens considerados fáceis.

A Função de Informação do Teste (FIT) é fornecida simplesmente pela soma das informações fornecidas por cada item que compõe o referido teste. Assim, a FIT é dada por:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta). \tag{2.6}$$

Outra forma de representar esta função de informação do teste é através do erro-padrão de medida, que em TRI é chamado de erro-padrão de estimação. Esta medida é calculada por:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}}. \tag{2.7}$$

É possível observar que, quanto maior for a informação do item, consequentemente a informação do teste, menor será o erro padrão de medida e, por conseguinte, maior a informação acerca de θ. Portanto, é importante ressaltar que essas medidas de informação dependem do valor de θ.

2.3.1.3 Suposições dos Modelos Dicotômicos

Entre várias características e pré-requisitos da TRI, há duas suposições básicas dos modelos propostos por essa técnica: unidimensionalidade e independência local.

A unidimensionalidade é a homogeneidade do conjunto de itens que teoricamente devem estar medindo um único traço latente, ou seja,

> um teste que é unidimensional consiste de itens que abrangem uma única dimensão. Sempre que apenas uma pontuação é descrita em um teste, não está implícito que os itens partilham de construto primário comum [...]. Unidimensionalidade, então, significa que um modelo possui um teta único para cada examinado e quaisquer outros fatores que afetam a resposta do item são considerados aleatórios ou erros de um único item, não afetando assim os demais itens (DEMARS, 2010, p.38).

Entretanto, é fácil supor que qualquer desempenho humano é sempre multimotivado ou multideterminado, visto que sempre mais do que um traço latente irá entrar na execução de qualquer que seja a tarefa. No entanto, segundo os autores Andrade et al. (2000), para satisfazer o postulado da unidimensionalidade, é suficiente admitir que haja uma habilidade dominante (um fator dominante) responsável

pelo conjunto de itens, sendo este o traço latente que se supõe estar sendo medido ou analisado pelo questionário.

Essa suposição pode ser verificada a partir da matriz de correlações tetracóricas através da análise fatorial. Outro procedimento sugerido na literatura é baseado no método de máxima verossimilhança.

Uma outra suposição do modelo em questão é conhecida como independência local ou independência condicional. Esse pressuposto afirma que dada uma habilidade medida pelo teste, as respostas aos diferentes itens do construto são independentes. Esta afirmação é fundamental para o processo de estimação dos parâmetros.

Conforme Demars (2010), se as respostas aos itens não são localmente independentes sob um modelo unidimensional, uma outra dimensão deve estar causando a dependência. Assim, de acordo com Andrade et al. (2000), a unidimensionalidade implica independência local. Consequentemente, o modelo tem então somente uma e não duas suposições a serem verificadas, de modo que os itens devem ser elaborados de modo a satisfazer somente a suposição de unidimensionalidade.

Além dessas duas principais suposições, é importante verificar também se o modelo está especificado corretamente, podendo assim o pesquisador usufruir das vantagens que a TRI oferece. Uma das vantagens adquiridas a partir de um bom ajuste, por exemplo, é a garantia de itens e habilidades invariantes.

2.3.2 Modelos para Itens Não Dicotômicos

Nesta seção, serão apresentados os modelos para itens não dicotômicos ou também chamados de modelos de respostas politômicas. Os modelos politômicos dependem da natureza das categorias de respostas. Nesses tipos de modelos, estão inclusos modelos tanto para análises de itens abertos, ou seja, de respostas livres, quanto para análise de itens de múltipla escolha, os quais são avaliados de forma graduada, compostos por itens que são elaborados ou corrigidos de modo a ter-se uma ou mais categorias intermediárias ordenadas entre as categorias certo ou errado.

Como neste tipo de item não se considera somente se o indivíduo respondeu corretamente ou não, mas também qual foi a alternativa escolhida por ele, utilizando assim mais intensamente a informação contida no teste, esses modelos necessitam de

um número maior de parâmetros a serem estimados. Serão apresentados a seguir alguns destes modelos.

2.3.2.1 Modelo de Resposta Nominal

O modelo de resposta nominal foi desenvolvido por Bock (1972), baseado no modelo logístico de dois parâmetros, com a característica adicional de poder ser aplicado a todas as categorias de respostas escolhidas em um teste com itens de múltipla escolha. Segundo Ferreira (2014), o objetivo do autor foi obter estimativas mais precisas dos traços latentes de indivíduos submetidos a testes de múltipla escolha, acrescentando, assim, a informação de um conhecimento parcial que o respondente apresenta relacionada à alternativa escolhida, sendo este negligenciado quando há a dicotomização das respostas.

Portanto, Bock (1972) definiu que a probabilidade de um indivíduo j selecionar uma particular opção k entre m_i opções avaliáveis do item i é representada por:

$$P_{i,k}(\theta_j) = \frac{e^{a^+_{i,k}(\theta_j - b^+_{i,k})}}{\sum_{h=1}^{m_i} e^{a^+_{i,h}(\theta_j - b^+_{i,h})}} \tag{2.8}$$

Com $i = 1, 2, ..., I$; $j = 1, 2, ..., n$ e $k = 1, 2, ..., m_i$. Em cada θ_j, a soma das probabilidades sobre as m_i opções, dada por $\sum_{k=1}^{m_i} P_{i,k}(\theta_j)$, é igual a 1. As quantidades $b^+_{i,k}$ e $a^+_{i,k}$ representam, respectivamente, os parâmetros relacionados à dificuldade e discriminação da *k-ésima* categoria do item *i*. Teoricamente, para esses valores são admissíveis quaisquer valores reais, e em particular para o parâmetro de discriminação, valores negativos são esperados para alternativas incorretas e positivos para as corretas. Já com relação ao parâmetro de dificuldade, espera-se que a opção correta apresente sempre o maior valor, pois essa opção exigirá um maior traço latente para ser escolhida. Este modelo assume que não há nenhuma ordenação a priori das opções de resposta. A Figura 2.6 contém uma representação gráfica deste modelo.

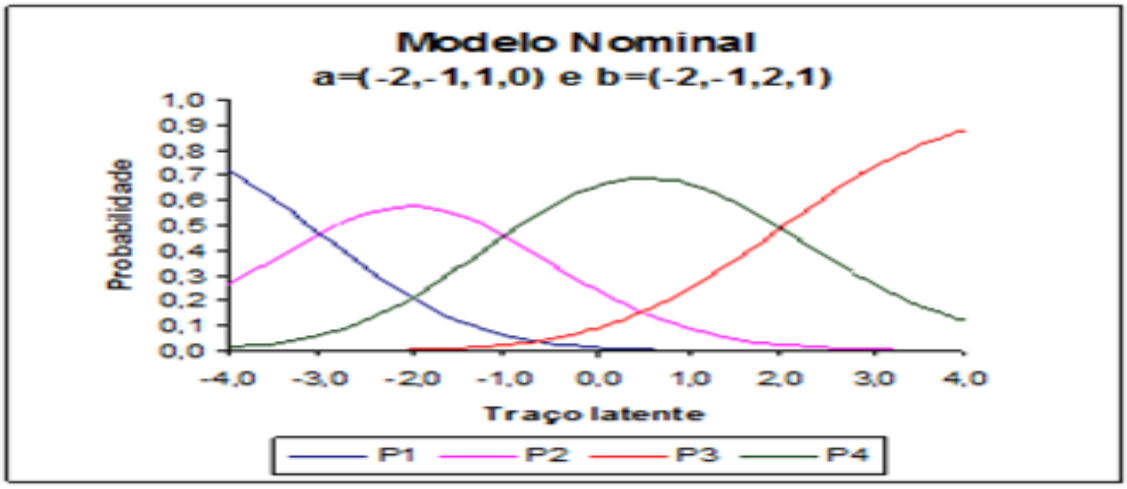

Figura 2.6: Representação Gráfica do Modelo de Resposta Nominal.
Fonte: Andrade (2005).

2.3.2.2 Modelo de Resposta Gradual

O modelo de resposta gradual de Samejima (1969) é uma generalização do modelo logístico de 2 parâmetros, assumindo que as categorias de resposta de um item podem ser ordenadas entre si, de tal forma que a categoria mais baixa contribua menos para o escore do indivíduo e a categoria mais alta contribua mais. A escala de Likert é um exemplo clássico desse tipo de modelo, em que as categorias poderiam ser definidas como: concordo completamente, concordo parcialmente, indiferente, discordo parcialmente e discordo completamente.

Assim como o modelo de Resposta Nominal, este modelo tenta obter mais informações das respostas fornecidas pelos respondentes, do que simplesmente se eles responderam corretamente ou não. Supondo que os escores das categorias de um item i dispostos em ordem do menor para o maior e denotados por $k = 1, 2, ..., m_i$ em que $(m_i + 1)$ é o número de categorias do *i-ésimo* item, de acordo com Andrade et al. (2000), a probabilidade de um indivíduo j escolher uma particular categoria ou outra mais alta do item i pode ser dada por:

$$P^{+}_{i,k}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k})}} \tag{2.9}$$

Com $i = 1, 2, ..., I$; $j = 1, 2, ..., n$ e $k = 0, 1, 2, ..., m_i$. O parâmetro $b_{i,k}$ representa o parâmetro de dificuldade da *k-ésima* categoria do item i e os demais

parâmetros são análogos aos já definidos na Seção 2.3.1. Diferentemente dos modelos dicotômicos, neste tipo de modelo o parâmetro de discriminação não pode ser chamado de parâmetro de inclinação, pois a discriminação de uma categoria específica de resposta não só depende do parâmetro de inclinação comum a todas as categorias do item, como também das distâncias das categorias de dificuldades adjacentes.

De acordo com a literatura, deve-se ter necessariamente uma ordenação entre o nível de dificuldade das categorias de determinado item i de acordo com a classificação de seus escores. Ou seja $b_{i,1} \leq b_{i,2} \leq ... \leq b_{i,m_i}$.

A probabilidade de um respondente j obter um escore k no item i é dada pela seguinte expressão:

$$P_{i,k}(\theta_j) = P^+_{i,k}(\theta_j) - P^+_{i,k+1}(\theta_j).$$

De modo que,

$$P^+_{i,0}(\theta_j) = 1,$$

e

$$P^+_{i,m_i+1}(\theta_j) = 0.$$

Logo, tem-se que

$$P_{i,0}(\theta_j) = P^+_{i,0}(\theta_j) - P^+_{i,1}(\theta_j) = 1 - P^+_{i,1}(\theta_j)$$

e

$$P_{i,m}(\theta_j) = P^+_{i,m}(\theta_j) - P^+_{i,m_i+1}(\theta_j) = P^+_{i,m}(\theta_j).$$

Portanto, temos que:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k})}} - \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k+1})}} \tag{2.10}$$

Percebe-se que, em um item com $(m_i + 1)$ opções de respostas, m_i valores de parâmetros de dificuldade necessitam ser estimados, além do parâmetro de inclinação do item. Dessa maneira, para cada item, o número de parâmetros a ser estimado será dado pelo seu número de categoria de respostas. Segue abaixo uma representação gráfica desse modelo (Figura 2.7).

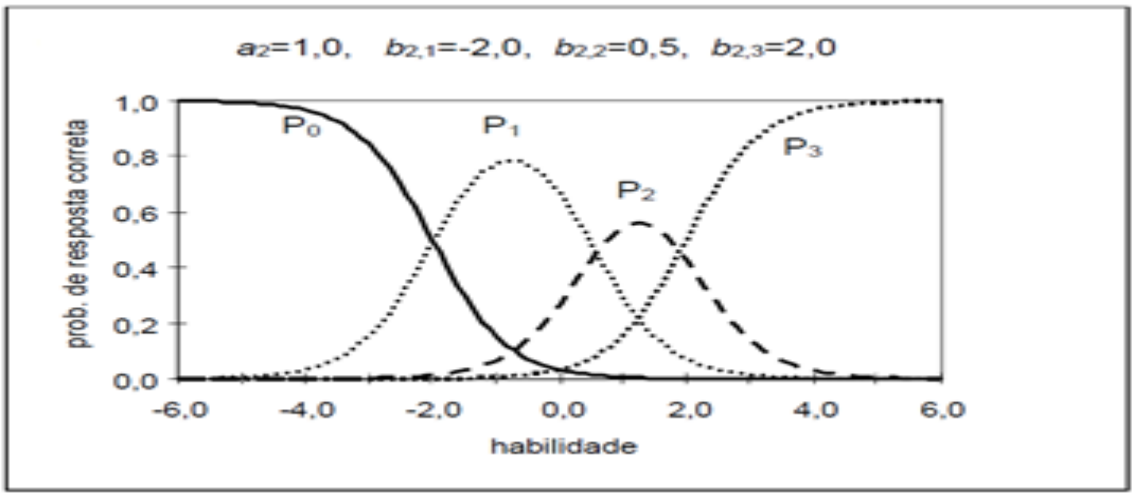

Figura 2.7: Representação Gráfica do Modelo de Resposta Gradual.
Fonte: Andrade, Tavares e Valle (2000).

2.3.2.3 Modelo de Escala Gradual

O modelo de escala gradual, proposto por Andrich(1978), é um caso particular do modelo de resposta gradual de Samejima (1969), sendo este também utilizado para itens com categorias de respostas ordenadas, porém com um pressuposto a mais: os escores das categorias devem ser igualmente espaçados. Este modelo possui *(m+1)* categorias representadas por *k=1,2,...,m*, ou seja, todos os itens têm as mesmas categorias de respostas. Este modelo é dado por:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i + d_k)}} - \frac{1}{1 + e^{-Da_i(\theta_j - b_i + d_{k+1})}} \tag{2.11}$$

Com $i = 1, 2, ..., I$; $j = 1, 2, ..., n$ e $k = 0, 1, 2, ..., m_i$.

Em que, b_i é agora o parâmetro de locação do item i e d_k é o parâmetro

de categoria. $P^{+}_{i,k}(\theta_j) - P^{+}_{i,k+1}(\theta_j) \geq 0$, então $d_k - d_{k+1} \geq 0$. Logo, devemos ter $d_1 \geq d_2 \geq ... \geq d_m$.

Segundo Alexandre et al (2003), b_{ik} é o parâmetro que representa a habilidade necessária para o *j-ésimo* indivíduo marcar a *k-ésima* categoria do *i-ésimo* item com uma certa probabilidade. Sendo assim, este parâmetro é medido na mesma escala da habilidade. Partindo do pressuposto de que no modelo de escala gradual os escores das categorias devem ser equidistantes, o parâmetro b_{ik} pode ser decomposto em um parâmetro b_i de locação do item e em um parâmetro d_k de categoria do item, ou seja, $b_{ik} = b_i - d_k$.

Ressalte-se que os parâmetros de categoria d_k não dependem do item, isto é, são comuns a todos os itens do instrumento de medição. Assim, se os itens que compõem o teste tiverem suas próprias categorias de respostas, podendo assim diferir no número, este modelo não será adequado.

Se um teste é composto por itens com $(m+1)$ categorias de resposta cada um, será necessário estimar m parâmetros de categoria, além dos parâmetros de inclinação e de locação de cada item. Logo, um modelo que tem I itens terá $2I + m$ parâmetros de itens a serem estimados. Na Figura 2.8, há uma representação gráfica do modelo de Escala Gradual.

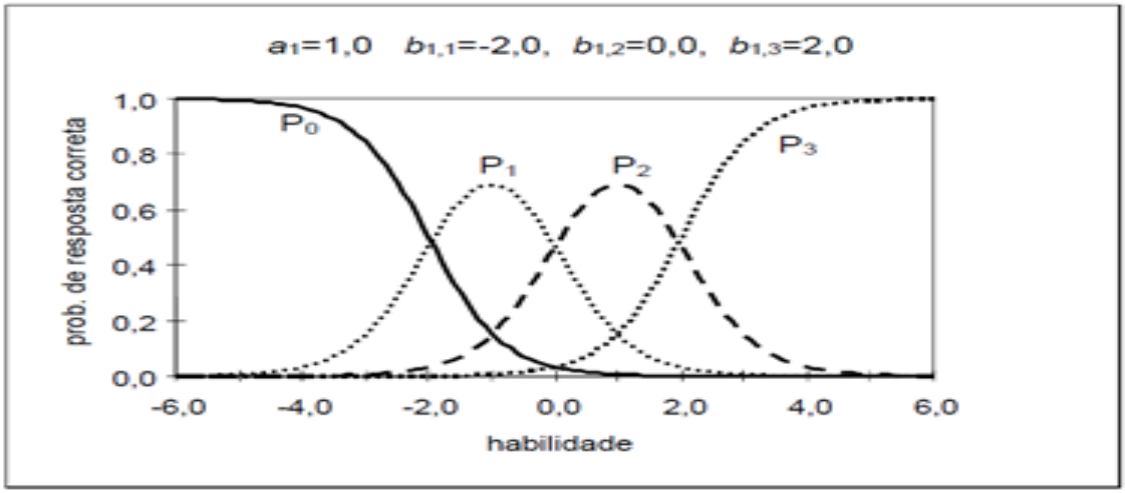

Figura 2.8: Representação Gráfica do Modelo de Escala Gradual.
Fonte: Andrade, Tavares e Valle (2000).

2.3.2.4 Modelo de Crédito Parcial

Outro modelo da TRI para respostas politômicas é o modelo de Crédito Parcial, desenvolvido por Masters (1982). Trata-se de uma extensão do modelo de Rasch para itens dicotômicos, o modelo logístico unidimensional de 1 parâmetro. Esse modelo é também utilizado na análise de respostas obtidas de duas ou mais categorias ordenadas, assim como gradual, e o que difere esses dois tipos de modelos é que o de crédito pertence à família de modelos de Rasch. Sendo assim, todos os parâmetros dos itens contidos no modelo são de locação, com parâmetro de discriminação comum a todos os itens.

Supondo que o item i tem $(m_i + 1)$ categorias de respostas ordenáveis $(k = 1, 2, ..., m_i)$, a probabilidade de um indivíduo com habilidade θ escolher a categoria k, dentre as $(m_i + 1)$ categorias do item i é dada por:

$$P_{i,k}(\theta_j) = \frac{\exp[\sum_{u=0}^{k}(\theta_j - b_{i,u})]}{\sum_{u=0}^{m_i}\exp[\sum_{v=0}^{u}(\theta_j - b_{i,v})]} \tag{2.12}$$

com $i = 1, 2, ..., I$; $j = 1, 2, ..., n$, $k = 0, 1, 2, ..., m_i$ e $b_{i,0} \equiv 0$,

em que $b_{i,k}$ é o parâmetro do item i que regula a probabilidade de escolher a categoria k em vez da categoria adjacente $(k - 1)$. De acordo com Andrade et al. (2000), cada parâmetro $b_{i,k}$ corresponde ao valor de habilidade em que o individuo tem a mesma probabilidade de responder à categoria k e à categoria $(k - 1)$, isto é, onde $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$.

Portanto, será necessário estimar m_i parâmetros de item para questões com $(m_i + 1)$ categorias de respostas. Percebe-se que modelos com itens de apenas duas opções de respostas se assemelharão com o modelo de Rasch para itens dicotômicos.

2.3.2.5 Modelo de Crédito Parcial Generalizado

O modelo de Crédito Parcial Generalizado, formulado por Muraki (1992), é baseado no modelo de crédito parcial de Masters, sendo que neste, a hipótese de poder de discriminação uniforme para todos os itens foi desconsiderada. A probabilidade de escolha da *k-ésima* categoria do item i, pelo indivíduo j, é dada por:

$$P_{i,k}(\theta_j) = \frac{\exp[\sum_{u=0}^{k} Da_i(\theta_j - b_{i,u})]}{\sum_{u=0}^{m_i} \exp[\sum_{v=0}^{u} Da_i(\theta_j - b_{i,v})]} \tag{2.13}$$

com $i = 1, 2, ..., I$; $j = 1, 2, ..., n$ e $k = 0, 1, 2, ..., m_i$. O número de categorias de respostas do item i é dado por $(m_i + 1)$, podendo este número variar de item para item. Portanto, somente m_i parâmetros de categoria do item podem ser identificados. Os parâmetros de categoria do item, $b_{i,k}$, são os pontos na escala de j em que as curvas $P_{i,k}(\theta_j)$ e $P_{i,k-1}(\theta_j)$ se interceptam, fato que acontece somente uma vez e em qualquer ponto da escala θ_j. Então, a partir da suposição de que a_i assume valores acima de 0, tem-se que:

- Se $\theta_j = b_{i,k}$, então $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$,
- Se $\theta_j > b_{i,k}$, então $P_{i,k}(\theta_j) > P_{i,k-1}(\theta_j)$,
- Se $\theta_j < b_{i,k}$, então $P_{i,k}(\theta_j) < P_{i,k-1}(\theta_j)$.

Da mesma forma como no modelo de escala gradual, no modelo de crédito parcial generalizado, o parâmetro $b_{i,k}$ pode ser decomposto através da diferença entre b_i e d_k.

O parâmetro d_k é interpretado como a dificuldade relativa da categoria k em comparação com as outras categorias do item, ou interpretado também como o desvio da dificuldade de cada categoria em relação à locação do item b_i. Assim, diferentemente do modelo de escala gradual, no modelo em questão os valores de d_k não são necessariamente ordenados sequencialmente dentro de uma questão. A Figura 2.9 ilustra graficamente o modelo de Crédito Parcial Generalizado.

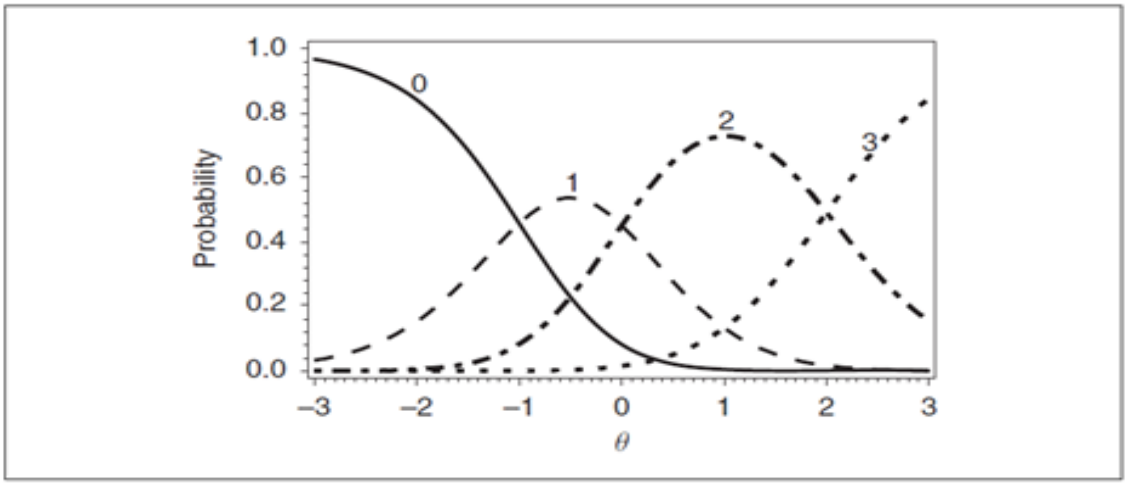

Figura 2.9: Representação Gráfica do Modelo de Crédito Parcial Generalizado. Fonte: Demars (2010).

2.4 Estimação dos Parâmetros da Teoria da Resposta ao Item

Um dos pontos críticos da TRI é a estimação dos parâmetros que caracterizam o modelo da resposta ao item: parâmetros dos itens e as habilidades dos respondentes. Nos modelos da TRI, como já mencionado anteriormente, a probabilidade de resposta correta a um determinado item depende exclusivamente da habilidade do examinado e dos parâmetros dos itens. Entretanto, na maioria das vezes, essas medidas são desconhecidas, sendo conhecidas somente as respostas aos itens dos testes. Assim, nos modelos da TRI tem-se um problema de estimação que envolve esses dois tipos de parâmetros.

Portanto, pode-se dividir o problema basicamente em três situações diferentes: Estimação das habilidades, quando já se conhecem os parâmetros dos itens; estimação dos parâmetros dos itens (processo conhecido como calibração) quando já se conhecem as habilidades e, por fim, a estimação conjunta das habilidades e dos parâmetros dos itens com, portanto, esses dois tipos de medidas desconhecidas.

Inicialmente, a estimação dos parâmetros em questão era feita através do método de máxima verossimilhança (MV) conjunta, o qual envolvia um grande número de parâmetros a serem estimados simultaneamente e gerava, consequentemente,

inúmeros problemas computacionais e teóricos, como a possível inconsistência dos estimadores obtidos por essa técnica. De acordo com Andrade et al. (2000), Bock e Lieberman, em 1970, desenvolveram um procedimento de estimação em duas etapas, denominado Método de Máxima Verossimilhança Marginal. Na primeira etapa, estimam-se os parâmetros dos itens, assumindo uma certa distribuição associada à habilidade dos indivíduos da população em estudo. Na segunda etapa, assumindo os parâmetros dos itens conhecidos, estimam-se assim as habilidades.

Apesar do avanço que esse método trouxe para a estimação dos parâmetros, ele continha algumas desvantagens matemáticas para as análises. Assim, em 1981, Bock e Aitkin propuseram que a obtenção das estimativas de máxima verossimilhança fossem feitas através da aplicação do algoritmo EM introduzido por Dempster, Laird e Rubin (1977). Mais recentemente, métodos bayesianos foram propostos também como uma opção de técnica para a estimação dos parâmetros.

Estimação das Habilidades

Na situação em que se deseja estimar somente as habilidades dos indivíduos, tem-se que os parâmetros dos itens são conhecidos. Utiliza-se esse tipo de estimação quando se deseja submeter indivíduos a itens já calibrados, visando à estimação de suas habilidades com o intuito de classificação ou seleção dos indivíduos. De acordo com Guewehr (2007), as habilidades dos respondentes podem ser estimadas a partir do método de máxima verossimilhança ou por métodos bayesianos, como o estimador bayesiano pela média da posteriori (EAP) ou o estimador pela moda da posteriori (MAP).

O método de máxima verossimilhança, ainda segundo a mesma autora, é um dos métodos mais populares de estimação por conter algumas boas propriedades como consistência das estimativas, conter erros padrões relativamente menores e com distribuição normal, produzir estimadores não viciados em testes "longos" (muitos itens), entre outras. Entretanto, o método de MV não é sustentado para alguns modelos por não conter determinados padrões de respostas.

O estimador bayesiano pela média da posteriori (EAP) é a média da distri-

buição a posteriori de θ e resolve alguns problemas do método anterior, pois consiste em incorporar qualquer informação a priori de modo a modificar a função de verossimilhança. Este estimador tem como vantagens o fato de poder ser definido para qualquer padrão de resposta e possuir menor erro médio do que qualquer outro estimador. No entanto, ele é um estimador viciado, exige cálculos mais complexos do que o método MV e necessita de uma distribuição a priori de θ.

O estimador pela moda a posteriori (MAP) consiste em obter o valor de θ que maximize a distribuição a posteriori. Este estimador apresenta características semelhantes ao EAP, pois é definido para qualquer padrão de resposta, porém também é viciado, exige cálculos mais complexos do que a MV e necessita de uma distribuição a priori de θ.

Estimação dos parâmetros dos itens

Quando se deseja estimar somente os parâmetros dos itens, necessita-se que as habilidades sejam conhecidas, o que na maioria das vezes na prática não ocorre. Um método comumente usado nessa situação é o de máxima verossimilhança marginal (MVM), o qual apresenta algumas vantagens em relação aos outros métodos. Dentre elas, pode-se citar o fato de possuir propriedades assintóticas, fazendo com que as estimativas dos parâmetros a, b e c sejam consistentes. Segundo Guewehr (2007), a ideia básica desse procedimento é a integração em θ, de modo que a função de verossimilhança não dependa dos parâmetros de habilidade. Nesse procedimento é necessário utilizar métodos iterativos e pode-se citar o comumente mais utilizado: o método de Newton-Raphson em sua forma multivariada.

Estimação dos parâmetros dos itens e das habilidades

A estimação dos parâmetros dos itens e das habilidades se trata do caso mais comum, em que nem os parâmetros dos itens são conhecidos e nem as habilidades. Então, estimam-se nessa situação os parâmetros dos itens e as habilidades ao mesmo tempo. Entretanto, devido à dificuldade de se estimar conjuntamente esse dois tipos de parâmetros, essa estimação é realizada em duas fases. Na primeira fase, é realizada a estimação dos parâmetros dos itens e esta pode ser feita tanto pelo método de

máxima verossimilhança conjunta (MVC), como por máxima verossimilhança marginal ou por métodos bayesianos. Em seguida, são estimadas as habilidades na mesma escala dos parâmetros dos itens, uma vez que eles foram estimados na fase anterior. Esta fase pode ser realizada por métodos de máxima verossimilhança ou por métodos bayesianos.

O método MVC, de acordo com Guewehr (2007), foi o primeiro método a ser utilizado nesse tipo de situação para a estimação dos parâmetros dos itens e serve como base para outros procedimentos. Por realizar uma solução simultânea das estimativas para todos os parâmetros, este método se torna bastante complexo. Esse procedimento de estimação apresenta algumas desvantagens, em que podemos citar a presença de problemas de indeterminação; não possuir propriedades assintóticas, pois quando o número de respondentes aumenta, o número de parâmetros a serem estimados também aumenta; ser bastante trabalhoso computacionalmente; não estar definido para alguns padrões de respostas; entre outras.

2.4.1 Métodos de Estimação

Nesta seção, trataremos de dois métodos de estimação comumente utilizados nas três diferentes situações definidas anteriormente: Quando se deseja estimar somente parâmetros dos itens, somente as habilidades ou estimar os parâmetros dos itens e as habilidades simultaneamente. Estes métodos serão abordados a seguir de forma resumida e para mais detalhes, consultar Andrade et al. (2000), Azevedo (2003), Baker (1992), Lord (1980) e Baker (2001).

Para auxiliar na apresentação destes métodos, além das definições apresentadas na seção 2.3.1, consideremos também as seguintes notações:

- $\boldsymbol{U}_{.j} = (U_{1j}, ..., U_{Ij})^{'}$, é o vetor de respostas do *j-ésimo* indivíduo aos I itens;
- $\boldsymbol{U}_{..} = (\boldsymbol{U}^{'}_{.1}, ..., \boldsymbol{U}^{'}_{.n})^{'}$, é o conjunto integral de respostas;
- $\boldsymbol{\zeta}_i = (a_i, b_i, c_i)^{'}$, é o vetor dos parâmetros do *i-ésimo* item;
- $\boldsymbol{\zeta} = (\boldsymbol{\zeta}^{'}_1, ..., \boldsymbol{\zeta}^{'}_I)^{'}$, é o vetor dos parâmetros de todos os itens;
- $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^{'}$, é o vetor de habilidades de todos os indivíduos.

De acordo com Azevedo (2003), duas principais suposições são necessárias para o desenvolvimento dos processos de estimação:

(S1) as respostas oriundas de indivíduos diferentes são independetes.

(S2) os itens são respondidos de forma independente por cada indivíduo (independência local) dada a sua habilidade.

2.4.1.1 Estimação por Máxima Verossimilhança

A seguir, será descrito o método de máxima verossimilhança para a estimação dos parâmetros dos itens, considerando as habilidades conhecidas e desconhecidas, e também a utilização do método para a estimação das habilidades.

Estimação dos Parâmetros dos Itens

Quando se deseja estimar os parâmetros dos itens na situação em que as habilidades são conhecidas, apesar desse cenário ser incomum, considerando as suposições (S1) e (S2), a função de verossimilhança $L(\boldsymbol{\zeta}) = P(\boldsymbol{U}_{\cdot\cdot} = \boldsymbol{u}_{\cdot\cdot}|\boldsymbol{\theta}, \boldsymbol{\zeta})$ pode ser dada por:

$$L(\boldsymbol{\zeta}) = \prod_{j=1}^{n} P(\boldsymbol{U}_{\cdot j} = \boldsymbol{u}_{\cdot j}|\theta_j, \boldsymbol{\zeta}) \tag{2.14}$$

$$= \prod_{j=1}^{n}\prod_{i=1}^{I} P(U_{ij} = u_{ij}|\theta_j, \zeta_i) \tag{2.15}$$

onde na última igualdade temos que a distribuição de U_{ij} só depende de $\boldsymbol{\zeta}$ através de ζ_i (pelo modelo). Usando a notação $P_{ij} = P(U_{ij} = 1|\theta_j, \zeta_i)$ e $Q_{ij} = 1 - P_{ij}$, temos que

$$P(U_{ij} = u_{ij}|\theta_j, \zeta_i) = P(U_{ij} = 1|\theta_j, \zeta_i)^{u_{ij}} P(U_{ij} = 0|\theta_j, \zeta_i)^{1-u_{ij}} \tag{2.16}$$

$$= P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \tag{2.17}$$

Portanto, a verossimilhança pode ser escrita como

$$L(\boldsymbol{\zeta}) = \prod_{j=1}^{n}\prod_{i=1}^{I} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \tag{2.18}$$

A log-verossimilhança, representada por $l(\boldsymbol{\zeta}) = lnL(\boldsymbol{\zeta})$, em que ln denota o logaritmo natural, pode ser escrita como:

$$l(\boldsymbol{\zeta}) = \sum_{j=1}^{n}\sum_{i=1}^{I}\{u_{ij} ln P_{ij} + (1-u_{ij}) ln Q_{ij}\} \tag{2.19}$$

De acordo com Andrade et al. (2000), os Estimadores de Máxima Verossimilhança (EMV) de ζ_i, $i = 1, 2, ..., I$, são os valores que maximizam a verossimilhança, ou, equivalentemente, são as soluções da seguinte equação:

$$\frac{\partial l(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} = 0, \ i = 1, 2, ..., I.$$

Dessa forma, o vetor escore (equações de estimação) resultante, segundo Azevedo (2003), é dado por:

$$\boldsymbol{S}(\boldsymbol{\zeta}_i) = \frac{\partial l(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} = \begin{bmatrix} D(1-c_i)\sum_{j=1}^{n}(u_{ij} - P_{ij})(\theta_j - b_i)W_{ij} \\ -Da_i(1-c_i)\sum_{j=1}^{n}(u_{ij} - P_{ij})W_{ij} \\ \sum_{j=1}^{n}\{(u_{ij} - P_{ij})\frac{W_{ij}}{P^*_{ij}}\} \end{bmatrix}$$

em que $W_{ij} = \frac{P^*_{ij}Q^*_{ij}}{P_{ij}Q_{ij}}$, $P^*_{ij} = \{1 + e^{-Da_i(\theta_j - b_i)}\}^{-1}$ e $Q^*_{ij} = 1 - P^*_{ij}$.

Como o sistema de equações descrito por $\boldsymbol{S}(\boldsymbol{\zeta}_i)$ não possui solução explícita, deve-se utilizar algum método iterativo para concluir a estimação. Os dois métodos mais usados neste tipo de situação são o Método de Newton-Raphson e Escore de Fisher. Para a aplicação destes métodos é necessário calcular a matriz Hessiana e a Informação de Fisher, dadas, respectivamente, por

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\zeta}_i) &= \sum_{j=1}^{n}\{(\frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}})(P^*_{ij}Q^*_{ij})\boldsymbol{H}_{ij} - (\frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}})^2(P^*_{ij}Q^*_{ij})^2\boldsymbol{h}_{ij}\boldsymbol{h}'_{ij}\} \\ &= \sum_{j=1}^{n}(u_{ij} - P_{ij})W_{ij}\{\boldsymbol{H}_{ij} - W_{ij}(u_{ij} - P_{ij})\boldsymbol{h}_{ij}\boldsymbol{h}'_{ij}\} \end{aligned} \tag{2.20}$$

e

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\zeta}_i) &= \mathbb{E}\{-\boldsymbol{H}(\boldsymbol{\zeta}_i)\} && (2.21)\\
&= -\sum_{j=1}^{n}\{W_{ij}\mathbb{E}(U_{ij}-P_{ij})\boldsymbol{H}_{ij} - W_{ij}^2\mathbb{E}(U_{ij}-P_{ij}^2)\boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^{'}\} && (2.22)\\
&= \sum_{j=1}^{n} P_{ij}Q_{ij}\frac{(P_{ij}^*Q_{ij}^*)^2}{(P_{ij}Q_{ij})^2}\boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^{'} = \sum_{j=1}^{n} P_{ij}^*Q_{ij}^*W_{ij}\boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^{'}, && (2.23)
\end{aligned}$$

em que

$$\boldsymbol{h}_{ij} = (P_{ij}^*Q_{ij}^*)^{-1}(\frac{\partial P_{ij}}{\partial \boldsymbol{\zeta}_i}) = \begin{bmatrix} D(1-c_i)(\theta_j - b_i) \\ -Da_i(1-c_i) \\ \frac{1}{P_{ij}^*} \end{bmatrix} \quad (2.24)$$

e

$$\begin{aligned}
\boldsymbol{H}_{ij} &= (P_{ij}^*Q_{ij}^*)^{-1}(\frac{\partial^2 P_{ij}}{\partial \boldsymbol{\zeta}_i \partial \boldsymbol{\zeta}_i^{'}})\\
&= \begin{bmatrix} D^2(1-c_i)(\theta_j-b_i)^2(1-2P_{ij}^*) & . & . \\ -D(1-c_i)\{1+Da_i(\theta_j-b_i)(1-2P_{ij}^*)\} & D^2a_i^2(1-c_i)(1-2P_{ij}^*) & . \\ -D(\theta_j-b_i) & Da_i & 0 \end{bmatrix}.
\end{aligned}$$

Assim, considerando $\widehat{\boldsymbol{\zeta}}_i^{(t)}$ uma estimativa de $\boldsymbol{\zeta}_i$ na iteração t, em que $t = 0, 1, 2, ...$, podemos definir os procedimentos de Newton-Raphson e o Escore de Fisher, respectivamente, como:

Newton-Raphson

$$\widehat{\boldsymbol{\zeta}}_i^{(t+1)} = \widehat{\boldsymbol{\zeta}}_i^{(t)} - [\boldsymbol{H}(\widehat{\boldsymbol{\zeta}}_i^{(t)})]^{-1}\boldsymbol{S}(\widehat{\boldsymbol{\zeta}}_i^{(t)}) \quad (2.25)$$

Escore de Fisher

$$\widehat{\boldsymbol{\zeta}}_i^{(t+1)} = \widehat{\boldsymbol{\zeta}}_i^{(t)} + [\boldsymbol{I}(\widehat{\boldsymbol{\zeta}}_i^{(t)})]^{-1}\boldsymbol{S}(\widehat{\boldsymbol{\zeta}}_i^{(t)}) \quad (2.26)$$

A literatura sugere alguns valores iniciais dos processos iterativos. Um procedimento alternativo para a estimação dos parâmetros dos itens quando as habilidades

são conhecidas é considerar um agrupamento dessas em classes. Para mais detalhes ver Andrade et al. (2000) e Azevedo (2003).

Por outro lado, quando se deseja estimar os parâmetros dos itens na situação em que as habilidades não são conhecidas, o método de estimação mais utilizado é o de Máxima Verossimilhança Marginal. A proposta desse método, de acordo com Andrade et al. (2000), é fazer a estimação em duas etapas: primeiro os parâmetros e em seguida, as habilidades. Como estas são desconhecidas, é necessário usar algum artifício de forma que a verossimilhança não seja função das habilidades. Assim, um artifício para eliminar as habilidades na verossimilhança consiste em marginalizar a verossimilhança, integrando-a com relação à distribuição da habilidade. De forma resumida, consideremos que as habilidades θ_j, $j = 1, ..., n$, são realizações de uma variável aleatória θ com distribuição contínua e função densidade de probabilidade $g(\theta|\boldsymbol{\eta})$ duplamente diferenciável, com componentes de $\boldsymbol{\eta}$ conhecidas e finitas. Para a situação em que θ tem distribuição Normal, temos que $\boldsymbol{\eta} = (\mu, \sigma^2)$, onde μ é a média e σ^2 é a variância das habilidades. Assim, pela suposição (S1), de acordo com Azevedo (2003), temos que a verossimilhança é dada por:

$$L(\boldsymbol{\zeta}, \boldsymbol{\eta}) \;\; = \;\; P(\boldsymbol{U}_{..}|\boldsymbol{\zeta}, \boldsymbol{\eta}) = \prod_{j=1}^{n} P(\boldsymbol{U}_{.j}|\boldsymbol{\zeta}, \boldsymbol{\eta}), \tag{2.27}$$

e, consequentemente, a logverossimilhança é dada por

$$l(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \sum_{j=1}^{n} lnP(\boldsymbol{U}_{.j}|\boldsymbol{\zeta}, \boldsymbol{\eta}). \tag{2.28}$$

Com os procedimentos descritos em Andrade et al. (2000), temos as seguintes equações de estimação:

$$a_i : (1 - c_i) \sum_{j=1}^{n} \int_{\mathbb{R}} [(u_{ij} - P_i)(\theta - b_i)W_i] g_j^*(\theta) = 0 \tag{2.29}$$

$$b_i : -a_i(1 - c_i) \sum_{j=1}^{n} \int_{\mathbb{R}} [(u_{ij} - P_i)W_i] g_j^*(\theta) = 0 \tag{2.30}$$

$$c_i : \sum_{j=1}^{n} \int_{\mathbb{R}} [(u_{ij} - P_i)\frac{W_i}{P_i^*}] g_j^*(\theta) = 0, \tag{2.31}$$

em que $W_i = \frac{P_i^* Q_i^*}{P_i Q_i}$, $P_i = c_i + (1 - c_i)\{1 + e^{-Da_i(\theta - b_i)}\}^{-1}$, $P_i^* = \{1 + e^{-Da_i(\theta - b_i)}\}^{-1}$, $Q_i = 1 - P_i$, $Q_i^* = 1 - P_i^*$ e $g_j^*(\theta) = \frac{P(\boldsymbol{U}_{.j}|\boldsymbol{\zeta},\theta) g(\theta|\boldsymbol{\eta})}{\int P(\boldsymbol{U}_{.j}|\boldsymbol{\zeta},\theta) g(\theta|\boldsymbol{\eta}) d\theta}$.

Uma vez que essas equações não possuem solução explícita, é necessário aplicar métodos com o intuito de resolvê-las. Para mais informações destes métodos, ver Andrade et al. (2000) e Azevedo (2003).

Estimação das Habilidade

No que diz respeito à estimação das habilidades (traços latentes) dos indivíduos, esta é feita de tal forma a considerar os parâmetros dos itens como conhecidos, estimados, por exemplo, pelos métodos citados anteriormente. Assim, neste caso, se é introduzido as estimativas dos parâmetros dos itens $\widehat{\boldsymbol{\zeta}}$ na verossimilhança original $L(\boldsymbol{\zeta}, \boldsymbol{\theta})$, obtendo $L(\boldsymbol{\theta}, \widehat{\boldsymbol{\zeta}})$. A partir disso, para o desenvolvimento desta seção, será considerado $L(\boldsymbol{\theta}, \widehat{\boldsymbol{\zeta}}) \equiv L(\boldsymbol{\theta})$. Então, baseado nas suposições (S1) e (S2), podemos escrever a log-verossimilhança como:

$$l(\boldsymbol{\theta}) = lnL(\boldsymbol{\theta}) \sum_{j=1}^{n} \sum_{i=1}^{I} \{u_{ij} lnP_{ij} + (1 - u_{ij}) lnQ_{ij}\} \tag{2.32}$$

De forma resumida, as expressões relativas aos processos de estimação, de acordo com Azevedo (2003), são dadas por:

Vetor Escore

$$S(\theta_j) = \sum_{i=1}^{I} a_i (1 - c_i)(u_{ij} - P_{ij}) W_{ij}, \tag{2.33}$$

Matriz Hessiana

$$H(\theta_j) = \sum_{i=1}^{I} (u_{ij} - P_{ij}) W_{ij} \{H_{ij} - (u_{ij} - P_{ij}) W_{ij} h_{ij}^2\}, \tag{2.34}$$

Informação de Fisher

$$I(\theta_j) = \sum_{i=1}^{I} P_{ij}^* Q_{ij}^* W_{ij} h_{ij}^2, \tag{2.35}$$

com

$$h_{ij} = (P_{ij}^* Q_{ij}^*)^{-1} (\frac{\partial P_{ij}}{\partial \theta_j}) = a_i(1 - c_i) \tag{2.36}$$

e

$$H_{ij} = (P_{ij}^* Q_{ij}^*)^{-1} (\frac{\partial^2 P_{ij}}{\partial \theta_j^2}) = a_i^2(1 - c_i)(1 - 2P_{ij}^*) \tag{2.37}$$

Então, considerando $\widehat{\theta_j}^{(t)}$ uma estimativa de θ_j na iteração t, em que t=0,1,2,..., temos que os métodos de Newton-Raphson e Escore de Fisher podem ser apresentados como:

Newton-Raphson

$$\widehat{\theta_j}^{(t+1)} = \widehat{\theta_j}^{(t)} - [H(\widehat{\theta_j}^{(t)})]^{-1} S(\widehat{\theta_j}^{(t)}) \tag{2.38}$$

Escore de Fisher

$$\widehat{\theta_j}^{(t+1)} = \widehat{\theta_j}^{(t)} + [I(\widehat{\theta_j}^{(t)})]^{-1} S(\widehat{\theta_j}^{(t)}) \tag{2.39}$$

2.4.1.2 Estimação Bayesiana

De acordo com Andrade et al. (2000), a estimação bayesiana consiste, basicamente, em estabelecer distribuições a priori para os parâmetros de interesse, construir uma nova função chamada de distribuição a posteriori e estimar os parâmetros de interesse baseado em alguma característica dessa distribuição.

A seguir, será descrito esse método de estimação bayesiana para a estimação dos parâmetros dos itens, considerando as habilidades conhecidas e desconhecidas, e será descrito também a utilização do método para a estimação das habilidades.

Estimação dos Parâmetros dos Itens

O processo de estimação bayesiana é baseado na construção da distribuição a posteriori, que une a informação contida na verossimilhança com a distribuição a priori. Dessa forma, a partir do Teorema de Bayes, de acordo com Azevedo (2003), na estimação dos parâmetros dos itens quando as habilidades são conhecidas a distribuição a posteriori é dada por:

$$
\begin{aligned}
f(\boldsymbol{\zeta}|\boldsymbol{u}_{..},\boldsymbol{\theta},\boldsymbol{\tau}) &= \frac{f(\boldsymbol{\zeta},\boldsymbol{U}_{..},\boldsymbol{\theta},\boldsymbol{\tau})}{\int_{\mathbb{R}^q} f(\boldsymbol{\zeta},\boldsymbol{U}_{..},\boldsymbol{\theta},\boldsymbol{\tau})d\boldsymbol{\zeta}} \propto f(\boldsymbol{U}_{..}|\boldsymbol{\theta},\boldsymbol{\zeta},\boldsymbol{\tau})f(\boldsymbol{\zeta},\boldsymbol{\tau}) && (2.40)\\
&= P(\boldsymbol{U}_{..}=\boldsymbol{u}_{..}|\boldsymbol{\theta},\boldsymbol{\zeta})f(\boldsymbol{\zeta}|\boldsymbol{\tau})f(\boldsymbol{\tau}) && (2.41)\\
&= \{\prod_{i=1}^{I}\prod_{j=1}^{n} P_{ij}^{u_{ij}}Q_{ij}^{1-u_{ij}}\}f(\boldsymbol{\zeta}|\boldsymbol{\tau})f(\boldsymbol{\tau}) && (2.42)
\end{aligned}
$$

com $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^{'},...,\boldsymbol{\tau}_I^{'})^{'}$, eventualmente $\boldsymbol{\tau}_i = (\boldsymbol{\tau}_{a_i}^{'},\boldsymbol{\tau}_{b_i}^{'},\boldsymbol{\tau}_{c_i}^{'})^{'} \equiv \boldsymbol{\tau}_i^{*} = (\boldsymbol{\tau}_{a_i}^{*'},\boldsymbol{\tau}_{b_i}^{*'},\boldsymbol{\tau}_{c_i}^{*'})^{'}$, representando os hiperparâmetros relacionados a $\boldsymbol{\zeta}_i$ e q o número de parâmetros dos itens. Considerando que os parâmetros de diferentes itens são estocasticamente independentes, ou seja, os parâmetros de determinado item não tem informações a respeito de outros, podemos reescrever a equação anterior, como:

$$
f(\boldsymbol{\zeta}|\boldsymbol{u}_{..},\boldsymbol{\theta},\boldsymbol{\tau}) \propto \{\prod_{i=1}^{I}\prod_{j=1}^{n} P_{ij}^{u_{ij}}Q_{ij}^{1-u_{ij}}\}\{\prod_{i=1}^{I} f(\boldsymbol{\zeta}_i|\boldsymbol{\tau}_i)\}\{\prod_{i=1}^{I} f(\boldsymbol{\tau}_i)\} \qquad (2.43)
$$

E o logaritmo natural da equação (2.43) é dado por:

$$
lnf(\boldsymbol{\zeta}|\boldsymbol{u}_{..},\boldsymbol{\theta},\boldsymbol{\tau}) \propto l(\boldsymbol{\zeta}) + \sum_{i=1}^{I} lnf(\boldsymbol{\zeta}_i|\boldsymbol{\tau}_i) + \sum_{i=1}^{I} lnf(\boldsymbol{\tau}_i) \qquad (2.44)
$$

com $l(\boldsymbol{\zeta}) = \sum_{i=1}^{I}\sum_{j=1}^{n}\{u_{ij}lnP_{ij} + (1-u_{ij})lnQ_{ij}\}$. Apesar de muitas opções de quantidades da distribuição a priori possam ser escolhidas como estimadores dos parâmetros, como média, moda, mediana, entre outras, será desenvolvido aqui, de forma resumida, o processo baseado na moda a posteriori (MAP). Então, como a distribuição a posteriori é contínua e dessa forma, a moda é o conjunto de valores que maximizam $f(\boldsymbol{\zeta}|\boldsymbol{u}_{..},\boldsymbol{\theta},\boldsymbol{\tau})$, temos que o vetor de funções de estimação bayesiana é dada por:

$$\boldsymbol{S}(\boldsymbol{\zeta}_i)_B = \frac{\partial lnf(\boldsymbol{\zeta}|\boldsymbol{u}_{..}, \boldsymbol{\theta}, \boldsymbol{\tau})}{\partial \boldsymbol{\zeta}_i} = \frac{\partial l(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} + \frac{\partial lnf(\boldsymbol{\zeta}_i|\boldsymbol{\tau}_i)}{\partial \boldsymbol{\zeta}_i}. \tag{2.45}$$

Nota-se que a primeira parcela do lado direito da equação anterior gera as equações de máxima verossimilhança $\boldsymbol{S}(\boldsymbol{\zeta}_i)$. Sendo assim, será desenvolvido aqui apenas a segunda parcela da equação (2.45), sendo esta referente às prioris dos parâmetros. Considerando que cada parâmetro a_i segue uma distribuição Log-normal com parâmetro $\boldsymbol{\tau}_{a_i} = (\mu_{a_i}, \sigma^2_{a_i})'$, cada b_i segue uma distribuição Normal $\boldsymbol{\tau}_{b_i} = (\mu_{b_i}, \sigma^2_{b_i})'$ e que cada parâmetro c_i segue uma distribuição Beta$(\alpha_i - 1, \beta_i - 1)$, temos então, de forma direta, que a segunda parcela do lado direito da equação (2.45) em relação a cada parâmetro do item a_i, b_i, c_i , respectivamente , é dadas por:

$$a_i : \frac{\partial lnf(a_i|\mu_{a_i}, \sigma^2_{a_i})}{\partial a_i} = -\frac{1}{a_i}[1 + \frac{lna_i - \mu_{a_i}}{\sigma^2_{a_i}}] \tag{2.46}$$

$$b_i : \frac{\partial lnf(b_i|\mu_{b_i}, \sigma^2_{b_i})}{\partial b_i} = -\frac{b_i - \mu_{b_i}}{\sigma^2_{b_i}} \tag{2.47}$$

$$c_i : \frac{\partial lnf(c_i|\alpha_i, \beta i)}{\partial c_i} = \frac{\alpha_i - 2}{c_i} - \frac{\beta_i - 2}{1 - ci}. \tag{2.48}$$

Então, a partir das equações apresentadas anteriormente, temos que o vetor de escores das Equações de Estimação Bayesianas (EEB) é dado por:

$$\boldsymbol{S}(\zeta_i)_B = \sum_{j=1}^{n}(u_{ij} - P_{ij})W_{ij}h_{ij} + \lambda_i, \tag{2.49}$$

com

$$\lambda_i = [\frac{1}{a_i}[1 + \frac{lna_i - \mu_{a_i}}{\sigma^2_{a_i}}]; -\frac{b_i - \mu_{b_i}}{\sigma^2_{b_i}}; \frac{\alpha_i - 2}{c_i} - \frac{\beta_i - 2}{1 - ci}]' \tag{2.50}$$

De acordo com Azevedo (2003), o sistema de equações dado por $\boldsymbol{S}(\zeta_i)_B$ é não linear, sendo necessário portanto, a utilização de algum método iterativo, como o de

Newton-Raphson ou Escore de Fisher. Para mais detalhes, ver Azevedo (2003), e Andrade (2000).

Na situação em que se deseja estimar os parâmetros dos itens em que as habilidades são desconhecidas, o método mais utilizado é chamado de Estimação Bayesiana Marginal (EBM). Dessa forma, de acordo com Azevedo (2003), a distribuição a posteriori é dada por:

$$\begin{aligned} f(\boldsymbol{\psi}|\boldsymbol{u}_{..}) &= C \iint L(\boldsymbol{u}_{..};\boldsymbol{\theta},\boldsymbol{\zeta}) f(\boldsymbol{\zeta}|\boldsymbol{\tau}) g(\boldsymbol{\theta}|\boldsymbol{\eta}) f(\boldsymbol{\tau}) g(\boldsymbol{\eta}) d\boldsymbol{\theta} d\boldsymbol{\tau} && (2.51)\\ &= Cg(\boldsymbol{\eta})\{\int L(\boldsymbol{u}_{..};\boldsymbol{\theta},\boldsymbol{\zeta}) g(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta}\}\{\int f(\boldsymbol{\zeta}|\boldsymbol{\tau}) f(\boldsymbol{\tau}) d\boldsymbol{\tau}\} && (2.52)\\ &\propto L(\boldsymbol{\zeta},\boldsymbol{\eta}) f(\boldsymbol{\zeta}) g(\boldsymbol{\eta}), && (2.53) \end{aligned}$$

Em que $\boldsymbol{\psi} = (\boldsymbol{\zeta}^{'}, \boldsymbol{\eta}^{'})^{'}$, $L(\boldsymbol{\zeta},\boldsymbol{\eta}) \equiv P(\boldsymbol{U}_{..}=\boldsymbol{u}_{..}|\boldsymbol{\zeta},\boldsymbol{\eta})$, $f(\boldsymbol{\zeta}) \equiv \int f(\boldsymbol{\zeta}|\boldsymbol{\tau}) f(\boldsymbol{\tau}) d\boldsymbol{\tau}$, C é uma constante de normalização, $L(\boldsymbol{u}_{..};\boldsymbol{\theta},\boldsymbol{\zeta})$ é a verossimilhança genuína, $f(\boldsymbol{\zeta}|\boldsymbol{\tau})$ e $g(\boldsymbol{\theta}|\boldsymbol{\eta})$ são as prioris associadas a $\boldsymbol{\zeta}$ e $\boldsymbol{\theta}$, respectivamente, $f(\boldsymbol{\tau})$ e $f(\boldsymbol{\eta})$ são as hiperprioris associadas a $\boldsymbol{\tau}$ e $\boldsymbol{\eta}$, respectivamente, em que $\boldsymbol{\tau}$ e $\boldsymbol{\eta}$, são os hiperparâmetros relativos a $\boldsymbol{\zeta}$ e $\boldsymbol{\theta}$, respectivamente.

O logaritmo natural dessa expressão é dado por:

$$\begin{aligned} lnf(\boldsymbol{\psi}|\boldsymbol{u}_{..}) \equiv l(\boldsymbol{\psi}|\boldsymbol{u}_{..}) &\propto lnL(\boldsymbol{\zeta},\boldsymbol{\eta}) + lnf(\boldsymbol{\zeta}) + lng(\boldsymbol{\eta}) && (2.54)\\ &\equiv l(\boldsymbol{\zeta},\boldsymbol{\eta}) + lnf(\boldsymbol{\zeta}) + lng(\boldsymbol{\eta}) && (2.55) \end{aligned}$$

Como o interesse é obter o máximo da posteriori (moda), deve-se derivar a função anterior com relação a ζ_i , resultando em:

$$\boldsymbol{S}(\zeta_i)_{BM} = \frac{\partial l(\boldsymbol{\zeta},\boldsymbol{\eta})}{\partial \zeta_i} + \frac{\partial lnf(\boldsymbol{\zeta})}{\partial \zeta_i} \qquad (2.56)$$

Novamente, se é necessário utilizar métodos interativos, como por exemplo o algoritmo EM, o qual é um processo interativo para a determinação de estimativas de máxima verossimilhança de parâmetros de modelos de probabilidade na presença de

variáveis aleatórias não observadas, podendo ser estendido ao caso em questão, de estimação bayesiana (moda a posteriori), ao substituir a esperança da log-posteriori (2.55) ao invés da esperança da verossimilhança.

Estimação das Habilidades

Equivalente ao que ocorre na estimação por máxima verossimilhança, a estimação bayesiana das habilidades é feita em uma segunda etapa considerando os parâmetros dos itens como fixos. A partir da suposição de independência entre as habilidades de diferentes indivíduos, ou seja, supondo que os traços latentes são estocasticamente independentes, a estimação para cada indivíduo pode ser feita separadamente. Assim, de acordo com Azevedo (2003), a distribuição a posteriori do traço latente de um determinado indivíduo é dada por:

$$g_j^*(\theta_j) \equiv Cg(\theta_j|\boldsymbol{u}_{.j}, \boldsymbol{\zeta}, \boldsymbol{\eta}) = CP(\boldsymbol{U}_{.j}|\theta_j, \boldsymbol{\zeta})g(\theta_j|\boldsymbol{\eta}) \quad (2.57)$$

$$\propto P(\boldsymbol{U}_{.j}|\theta_j, \boldsymbol{\zeta})g(\theta_j|\boldsymbol{\eta}). \quad (2.58)$$

Segundo Andrade et al. (2000), podemos adotar alguma característica de $g_j^*(\theta_j)$ como estimador da habilidade θ_j, em que as mais adotadas são a média [conhecida por estimação pela média da posteriori (EAP - Expectation a posteriori)] e a moda [conhecida por estimação pela moda da posteriori (MAP - Maximum a posteriori)].

Na estimação pela moda da posteriori ou MAP, devemos encontrar o máximo da distribuição a posteriori (2.58) e por conveniência, devemos tomar o logaritmo natural desta equação. Assim, de acordo com Azevedo(2003), este é dado por:

$$lng_j^*(\theta_j) \equiv l_j^*(\theta_j) \propto l(\theta_j) + lng(\theta_j|\boldsymbol{\eta}) \quad (2.59)$$

com $l(\theta_j) = \prod_{i=1}^{I} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$. Dessa forma, a função de estimação bayesiana, para determinada habilidade, é dada por:

$$\boldsymbol{S}(\theta_j)_B = \frac{\partial l_j^*(\theta_j)}{\partial \theta_j} = \frac{\partial l(\theta_j)}{\partial \theta_j} + \frac{\partial lng(\theta_j|\boldsymbol{\eta})}{\partial \theta_j}. \tag{2.60}$$

Assumindo que as habilidades seguem uma distribuição normal com $\eta = (\mu_\theta, \sigma_\theta^2)'$ temos que:

$$\boldsymbol{S}(\theta_j)_B = \sum_{j=1}^{I} a_i(1-c_i)(u_{ij} - P_{ij}) - \frac{\theta_j - \mu_\theta}{\sigma_\theta^2}). \tag{2.61}$$

Novamente, pela não linearidade da equação de estimação, devemos usar algum método iterativo para resolvê-la, como o método "Scoring" de Fisher. Para mais detalhes, consultar Andrade et al. (2000).

Na estimação pela média da posteriori ou EAP, de acordo com Andrade et al. (2000), a posteriori é dada por:

$$g(\theta|\boldsymbol{u}_{.j}, \boldsymbol{\zeta}, \boldsymbol{\eta}) = \frac{P(\boldsymbol{u}_{.j}|\theta, \boldsymbol{\zeta}) g(\theta|\boldsymbol{\eta})}{P(\boldsymbol{u}_{.j}|\boldsymbol{\zeta}, \boldsymbol{\eta})}. \tag{2.62}$$

Baseado no objetivo da EAP, o qual é obter a esperança da posteriori, esta é dada por:

$$\widehat{\theta}_j \equiv E[\theta|\boldsymbol{u}_{.j}, \boldsymbol{\zeta}, \boldsymbol{\eta}] = \frac{\int_{\mathbb{R}} \theta g(\theta|\boldsymbol{\eta}) P(\boldsymbol{u}_{.j}|\theta, \boldsymbol{\zeta}) d\theta}{\int_{\mathbb{R}} g(\theta|\boldsymbol{\eta}) P(\boldsymbol{u}_{.j}|\theta, \boldsymbol{\zeta}) d\theta} \tag{2.63}$$

Dessa forma, esta estimação tem a vantagem de ser calculada diretamente, não necessitando de processos iterativos e, além disso, as quantidades necessárias para o seu cálculo são um produto final da fase da estimação, afirmam Andrade et al. (2000).

2.5 Contextualização da Teoria da Resposta ao Item neste estudo

Neste trabalho, será simulado um banco de dados em que iremos supor sua origem a partir da aplicação de uma prova de conhecimentos na qual foi medida alguma habilidade. Este banco de dados será considerado incompleto, visto que será caracterizado por questões respondidas e questões não respondidas. Logo, será um banco de dados com informações faltantes, as quais serão descritas e discutidas no próximo capítulo.

Iremos supor diferentes motivos para a presença de informações faltantes, tais como: falta de tempo para resolver a prova, falta de conhecimento do aluno sobre o conhecimento medido, prova cansativa, entre outros. Além destes motivos, iremos considerar ainda, a não resposta como uma escolha consciente do aluno em deixar a questão em branco, caso ele não tenha certeza da resposta correta. Essa suposição será baseada nas regras de correção das provas objetivas do vestibular da UnB (Universidade de Brasília), elaboradas pelo CEBRASPE (Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos).

As provas objetivas deste teste são caracterizadas por quatro tipos de questões, denominados tipo A, tipo B, tipo C e tipo D. De forma resumida, as questões são definidas da seguinte forma: As questões do tipo A têm como opções de respostas "Certo" ou "Errado". Nas questões do tipo B, é proposto um problema ao candidato e ele deve marcar um único resultado numérico como resposta da questão, representado por um número inteiro de 000 a 999. As questões do tipo C têm quatro opções de respostas, designadas pelas letras A, B, C e D, das quais apenas uma constitui a resposta correta. Por fim, as questões do tipo D são itens de respostas construídas, ou seja, questões abertas, com respostas elaboradas pelo candidato.

Neste estudo, consideraremos que o banco de dados será constituído com questões do tipo A as quais, além de apresentarem as opções "certo" ou "errado", como citado anteriormente, são corrigidas a partir do seguinte cálculo: caso a resposta do candidato esteja em concordância com o gabarito oficial definido na prova, ou seja, caso ele acerte a questão, ele tem uma pontuação +1 (um ponto positivo). Caso a

resposta do candidato esteja em discordância com o gabarito oficial definido na prova, ou seja, caso ele erre a questão, ele tem uma pontuação -1 (um ponto negativo). E, por fim, caso não haja marcação por parte do candidato, ele tem pontuação 0 (zero pontos). Estas informações foram extraídas do Edital Nº1, do 1º vestibular da UnB de 2014, lançado no dia 22 de abril de 2014.

Considerando o cálculo, é constatado que cada resposta errada marcada pelo candidato anulará uma resposta certa que ele já tenha adquirido ou venha a adquirir. Percebe-se que em casos de incertezas do candidato quanto à resposta correta da questão, a não resposta costuma ser vantajosa. Sendo assim, no banco de dados utilizado nesta pesquisa, as respostas faltantes serão consideradas também como uma opção de resposta, ou seja, teremos três opções de respostas: "Errar", "Não Responder" e "Acertar".

A partir da suposição considerada acima, o modelo utilizado neste trabalho será o modelo da teoria da resposta ao item, com somente uma população envolvida no estudo; unidimensional, ou seja, com apenas um traço latente medido e de natureza dos itens não dicotômica ou também chamada de "politômica". O modelo para itens não dicotômicos considerado será o modelo de Resposta Gradual, em que as categorias de resposta de uma questão podem ser ordenadas entre si, de tal forma que a categoria mais baixa contribua menos para o escore do respondente e a categoria mais alta contribua mais. A principal razão para a escolha desse modelo para ser utilizado neste trabalho é exatamente a característica de se ter uma ordenação entre as categorias, visto que isso garantirá que a categoria de "não resposta" será uma categoria intermediária entre as outras categorias de resposta: "acertar" e "errar". Dessa forma, verificamos que os outros modelos para itens não dicotômicos da TRI não são adequados a nossa proposta, como por exemplo, o Modelo de Resposta Nominal, o qual assume que não há nenhuma ordenação entre as categorias. Com relação ao Modelo de Escala Gradual, que, apesar de ser um caso particular do modelo de resposta gradual, tem uma característica adicional que não necessariamente seria uma exigência da nossa proposta: os escores das categorias são igualmente espaçados. Por fim, os modelos de crédito parcial e crédito parcial generalizado também não foram escolhidos por se basearem no modelo logístico de 1 parâmetro, tendo somente o parâmetro de dificuldade, sendo o parâmetro de discriminação comum a todos os itens. Dessa

forma, como também temos o intuito de avaliar o quão bem os itens discriminam os indivíduos, esses dois últimos modelos não estavam também de acordo com a nossa proposta. Vale ressaltar novamente, que as categorias consideradas serão: “Errar”, “Não Responder” e “Acertar”.

Capítulo 3

Dados Faltantes

3.1 Introdução

Na maioria das pesquisas em que se trabalha com banco de dados, é comum deparar-se com dados faltantes ou também chamados dados perdidos ou *missings*, gerando-se assim, banco de dados incompletos. Segundo Mcknight et al. (2007), a expressão dados faltantes significa, em termos gerais, a perda de algum tipo de informação sobre o fenômeno em que estamos interessados.

Esse tipo de problema tem se tornado cada vez mais frequente em diferentes áreas de pesquisas científicas como ciências sociais, educação, saúde, entre outras. Pesquisadores podem deparar com dados faltantes em seus bancos de dados por diferentes motivos, os quais podem surgir tanto na etapa da realização da pesquisa, como na etapa da coleta de dados. Alguns desses motivos são: não preenchimento cadastral; falta de cooperação do entrevistado, pois alguns se recusam a responder ou são incapazes de dar a resposta correta a um ou mais itens por falta de conhecimento no assunto; pesquisas com questionários ou testes muito longos e cansativos, cuja extensão faz o respondente não realizar o teste até o fim; curto espaço de tempo para responder os quesitos; questões ou itens mal formulados; falha do entrevistador ao perguntar ou registrar a resposta; erro de digitação ao se fazer o registro dos dados; possíveis problemas no armazenamento dos dados, entre outros.

A ocorrência de dados faltantes é uma limitação bastante delicada, tornando

um desafio o uso do banco de dados incompletos. Isso se dá, entre outros motivos, pelo fato de que a maioria das técnicas estatísticas são desenvolvidas para serem utilizadas em matrizes de dados completas. Portanto, ao se utilizarem técnicas estatísticas não adequadas para banco de dados incompletos, geram-se conclusões errôneas sobre as informações estudadas e, por consequência, a perda da eficiência das estimativas e o surgimento de vieses pelo fato de, por exemplo, frequentemente existir diferenças de respostas entre respondentes e não respondentes.

Segundo Farhangfar et al. (2007, apud PEREIRA, 2014), três principais tipos de problemas estão associados à presença de dados faltantes: perda de eficiência; complicações na manipulação e na análise de dados; e viés, resultantes das discrepâncias entre os valores atribuídos aos dados faltantes e os valores reais desconhecidos. Portanto, a perda de dados pode comprometer a qualidade dos resultados produzidos, neles interferindo e conduzindo a interpretações indevidas.

Estratégias para evitar e lidar com dados faltantes em variáveis importantes da pesquisa devem ser definidas no decorrer da fase de planejamento do estudo, na coleta de dados ou no tratamento dos dados faltantes com métodos estatísticos apropriados e elaborados para resolver esse tipo de problema. Alguns autores sugerem diferentes formas de prevenção de perdas de informação, sendo estas medidas necessariamente realizadas antes e durante a aplicação do teste ou questionários. Podemos citar como estratégias, por exemplo, o uso de incentivos para estimular o indivíduo a responder todo o teste, analisar o melhor modo de aplicar o questionário de acordo com o público alvo, igualar a etnia e idade do entrevistador e o entrevistado, entre outras. De acordo com Mcknight et al. (2007), diminuir a responsabilidade do entrevistado de responder o teste e aumentar os benefícios que ele possa vir a adquirir por participar da pesquisa faz com que haja uma diminuição na incidência de dados faltantes.

No entanto, a prevenção nem sempre é possível e, pelos motivos citados anteriormente, há uma real necessidade de tratar os dados faltantes ao deparar com eles no banco de dados em estudo, ao invés de simplesmente excluí-los ou ignorá-los. Baseado nisso, o desenvolvimento de técnicas estatísticas direcionadas para a resolução dos problemas gerados por dados faltantes tem sido uma área de pesquisa bastante ativa nas últimas décadas.

Essas técnicas são denominadas “Imputação de Dados Faltantes” e têm por

objetivo completar os bancos de dados envolvendo a substituição dos dados faltantes por estimativas de valores plausíveis a serem imputados no lugar dos respectivos dados faltantes. Dessa forma, essas técnicas completam as bases de dados, possibilitando assim a análise com todas as informações em estudo.

De acordo com Assunção (2012), o método mais simples, e que também está disponível na maioria dos softwares estatísticos, é a substituição dos dados faltantes por alguma medida resumo, em que os dados faltantes são substituídos pela média ou mediana dos dados válidos, sendo estes considerados " dados não faltantes". Apesar deste método ser de implantação fácil e imediata, e por consequência, bastante utilizado, ele resulta em algumas desvantagens, como a introdução artificial de uma baixa estimativa da variabilidade da variável, provocando, por exemplo, a obtenção de intervalos de confiança inadequados ou viciados. Outra desvantagem, comumente citada na literatura, é a diminuição da relação com as demais variáveis, o que impossibilita, por exemplo, a utilização de outras variáveis do próprio conjunto de dados para aprimorar o processo de imputação.

A imputação única, por sua vez, consiste em outra técnica também frequentemente utilizada, em que o dado faltante é substituído por valores previstos gerados a partir dos " dados não faltantes" das demais variáveis contidas no banco de dados, gerando assim estimativas mais consistentes dos dados faltantes. Para produzir tais estimativas são utilizadas técnicas estatísticas, tais como: regressão linear, regressão multinomial, algoritmos EM, entre outras.

Por outro lado, Nunes (2007) afirma que, apesar de essa técnica preencher os dados faltantes, obtendo-se assim um banco de dados completos para ser utilizado na análise, a incerteza associada à imputação deve ser levada em conta para que os resultados obtidos com os dados completos sejam válidos, pois os valores imputados não são valores reais. Em outras palavras, de qualquer forma, sempre haverá um erro gerado por este processo de estimação e essa incerteza precisa ser levada em consideração durante a análise dos resultados gerados através da base completada pela imputação.

Com a intenção de resolver essa questão, Rubin (1987) desenvolveu a técnica de Imputação Múltipla, que, de forma simplificada, consiste em executar alguns dos processos de imputação citados anteriormente repetidas vezes, gerando-se, assim,

múltiplos bancos de dados imputados. A análise estatística escolhida para a análise dos dados é então realizada em cada um destes "novos" bancos de dados, produzindo-se diferentes resultados, sendo estes combinados de forma tal que se gere assim, um resultado final. A quantidade de combinações a serem realizadas é definida de acordo com o estudo em questão.

A imputação de dados faltantes é a prática mais comum encontrada na literatura em trabalhos em que os pesquisadores se deparam com dados faltantes em seus bancos de dados. Porém, além desta prática, existem outras formas de tratamento de dados faltantes que também têm apresentado um desempenho superior em relação aos métodos mais "tradicionais", de acordo com Assunção (2012). Assim como os métodos mais avançados de imputação, estes outros métodos alternativos não se concentram somente em identificar um substituto para o valor faltante, mas também em levar em consideração todas as características do banco de dados, com o intuito de aproveitar a maioria de informações disponíveis, preservando assim as relações existentes no conjunto do banco de dados.

3.2 Tipos de Dados Faltantes

Como a perda de informações é praticamente inevitável, um aspecto importante durante a análise de dados, antes mesmo que seja aplicado qualquer método de tratamento para a resolução desse problema, é a identificação dos mecanismos que causaram os dados faltantes, ou seja, o motivo pelo qual surgiram os dados faltantes na base de dados. Nunes (2007) afirma que é importante que os dados faltantes não sejam considerados apenas como um problema de análise de dados, mas também como uma questão de planejamento da pesquisa e de interpretação dos dados, sendo necessária também, a identificação dos mecanismos que geram os dados faltantes.

Baseado nesta questão, surgiram diferentes mecanismos de não respostas. De acordo com Mcknight et al. (2007), o sistema de classificação mais amplamente utilizado foi introduzido por Donald Rubin (1976), o qual especificou três categorias distintas de dados faltantes: a) Dados faltantes completamente aleatórios (MCAR), b) Dados faltantes aleatórios (MAR) e c) Dados faltantes não aleatórios (MNAR). Os autores afirmam ainda que os termos citados anteriormente referem-se à proba-

bilidade de dados faltantes, levando em consideração as informações de basicamente três aspectos: variáveis que contenham dados faltantes, variáveis associadas aos dados faltantes (covariáveis) e um mecanismo hipotético subjacente aos dados faltantes. A seguir serão apresentadas as características da classificação de Donald Rubin para este tipo de dado.

3.2.1 Dados Faltantes Completamente Aleatórios

Os dados faltantes são considerados MCAR (*Missing Completely at Random*) quando as razões para as perdas das informações não estão relacionadas a quaisquer respostas dos indivíduos, incluindo o dado em falta. Em outras palavras, podem-se classificar os dados faltantes como MCAR quando a probabilidade de uma questão ter respostas ausentes não depende nem dos valores observados e nem dos valores não observados.

Isso significa que, de acordo com Veroneze (2011), a causa que levou aos dados faltantes é um evento aleatório e esta causa não está relacionada com os dados. Assim, os valores faltantes para uma variável aleatória são uma simples amostra aleatória dos dados dessa variável, significando que a distribuição dos valores faltantes é de mesma natureza da dos valores observados.

Segundo Acock (2005), o termo MCAR tem um significado preciso: ao pensarmos no conjunto de dados como uma grande matriz, pode-se afirmar que os dados faltantes estarão distribuídos aleatoriamente por toda a matriz. O autor afirma ainda que a única limitação desta classe de dados faltantes é que a incerteza é introduzida pelo processo de imputação, e essa incerteza reduz o poder estatístico ao se comparar com os dados completos.

Uma das grandes vantagens do dado faltante ser MCAR é que a causa que gerou os dados faltantes não precisa ser levada em consideração na análise para controlar a influência destes nos resultados da pesquisa, de acordo com Veroneze (2011). Entretanto, afirmar que os dados são MCAR pode ser uma hipótese forte e pouco realista na prática, uma vez que na maioria dos experimentos, geralmente existe um grau de relação entre os dados faltantes e as informações das covariáveis.

3.2.2 Dados Faltantes Aleatórios

Os dados ausentes são classificados como MAR (*Missing at Random*) quando as respostas faltantes dependem somente das variáveis observadas e não mais dos dados ausentes, podendo assim, ser explicadas pelas demais variáveis presentes no banco de dados. Dito de outra forma, dados faltantes MAR ocorrem quando o padrão de perda em uma variável pode ser predito a partir de outras variáveis do banco de dados, não sendo, portanto, devido à variável específica na qual os dados são ausentes.

De acordo com Acock (2005), a suposição de que os dados faltantes são MAR só é válida se pudermos assumir que o padrão dos dados faltantes é condicionalmente aleatório, dados os valores observados das outras variáveis. Veroneze (2011) afirma que, neste caso, os dados faltantes de uma variável são como uma amostra aleatória simples das informações para essa variável dentro de subgrupos definidos pelos valores observados, e, desta maneira, os valores em falta possuem a mesma distribuição dos valores observados dentro de cada subgrupo.

Uma das vantagens de haver dados faltantes dependentes das variáveis preenchidas é que estes podem ser explicados pelas demais variáveis presentes no banco de dados e ao realizar-se o tratamento desses elementos, é possível obter uma análise não viesada, considerando as informações que "causam" os dados faltantes.

3.2.3 Dados Faltantes Não Aleatórios

De acordo com Acock (2005), os dados faltantes podem não ser classificados nem como MCAR nem como MAR, mas, no entanto, serem sistemáticos. Estes são classificados então como MNAR (*Missing Not at Random*) e ocorrem quando são gerados de forma não mensurável, isto é, eles dependem de eventos que o pesquisador não consegue observar e controlar. Sendo assim, pode-se afirmar que os itens faltantes dependem dos dados não observados e, em algumas vezes, de acordo com Veroneze (2011), podem depender também dos dados observados.

Nunes (2007) afirma que dados que são mais prováveis de serem faltantes, geralmente são aqueles situados nas posições extremas da distribuição. Esses valores, por sua vez, são mais altos ou mais baixos do que o padrão da amostra estudada.

Resumindo as principais diferenças entre as classes dos mecanismos dos dados faltantes, temos: o mecanismo é dito MCAR se os dados faltantes são originados por processos aleatórios. Já o mecanismo é declarado MAR se os dados faltantes são causados por uma ou mais variáveis observadas. Por último, o mecanismo é dito MNAR quando os dados faltantes são causados por uma ou mais variáveis não-observadas, podendo existir também, uma relação entre os dados observados e não observados, neste último mecanismo (VERONEZE, 2011).

De acordo com McKnight et al.(2007), há outra distinção pertinente no sistema de classificação de dados faltantes de Rubin. Os mecanismos podem ainda ser divididos em *ignoráveis* ou *não ignoráveis*, em que os dados MCAR e MAR se encaixam na primeira categoria e os dados NMAR na segunda, respectivamente. O autor afirma ainda que o termo "ignorável", no contexto de mecanismos de dados faltantes, não significa que os pesquisadores devam simplesmente ignorar este tipo de dado, mas sim considerar este tipo de mecanismo como parte do processo de estimação dos parâmetros.

Segundo ainda o mesmo autor, os mecanismos classificados como ignoráveis são considerados mais fáceis de tratar, uma vez que seus efeitos nos modelos estatísticos estão disponíveis para o pesquisador. Assim, o mecanismo MCAR não deve gerar grandes impactos na estimação dos parâmetros, visto que as informações faltantes deste tipo são geradas de forma completamente aleatória. Da mesma forma, para o mecanismo MAR, há um processo sistemático subjacente à falta de dados aleatória, o qual pode ser modelado a partir dos dados observados da matriz.

Por outro lado, quando o mecanismo é "não ignorável", e portanto não há informações dentro do conjunto de dados que possibilite a modelagem e a compreensão da maneira como os dados faltantes aconteceram, o efeito do mecanismo é desconhecido e potencialmente perigoso.

Além dos autores McKnight et al.(2007), Graham (2009) também discute a classificação de Rubin e comenta ainda, suas principais consequências. Segundo o autor, o principal efeito dos dados MCAR é a perda do poder estatístico, produzindo parâmetros imparciais, assim como os dados MAR. Já os dados MNAR são considerados um problema, visto que eles produzem estimativas dos parâmetros tendenciosas.

De maneira geral, é possível dividir os motivos de produção de dados faltantes

em aqueles que podem ser explicados (MCAR e MAR) e os que não podem (MNAR). De forma prática, a maioria dos dados faltantes têm os motivos de sua criação caracterizados por fatores que podem e que não podem ser explicados. Entretanto, a escolha de métodos apropriados para o tratamento desses dados faz com que esses dados possam expressar bons resultados, mesmo em circunstâncias em que o analista não tenha certeza se os dados foram causados por motivos somente mensuráveis.

3.3 Métodos de Exclusão

Como foi já foi citado anteriormente, não dar a devida atenção aos dados faltantes de um banco de dados em estudo pode gerar sérios problemas nos resultados das análises. Por esse motivo, além dos métodos de imputação apresentados anteriormente, existem outros métodos de tratamento dos dados faltantes, que são, inclusive, mais frequentemente utilizados na prática. Estes métodos são chamados de Métodos de Exclusão, os quais podem ser divididos em dois principais tipos: Método de Exclusão *Listwise* e Método de Exclusão *Pairwise*. Esses métodos serão apresentados a seguir.

3.3.1 Método de Exclusão *Listwise*

O método *listwise*, ou também conhecido pelo método de Análises de Casos Completos, consiste em excluir todos os casos que contenham uma ou mais informações faltantes em qualquer variável, ou seja, qualquer indivíduo ou participante que tenha valores ausentes é simplesmente descartado do banco de dados, deixando-se apenas os casos com informações completas para todas as variáveis. Vários autores, como Acock (2005), Little (1992), Mcknight at al. (2007), Pigott (2001) e Van Buuren (2012), afirmam em seus estudos que o *listwise* é o método de exclusão mais comum utilizado nas análises, sendo este o procedimento padrão ou o "default" da maioria dos softwares estatísticos, tais como os softwares R, SPSS, SAS e Stata.

Existem várias vantagens oriundas da aplicação deste método. Dentre elas, podemos citar a conveniência e a simplicidade de como a técnica é executada. De acordo com Pigott (2001), a principal vantagem deste método é a facilidade de imple-

mentação, uma vez que o pesquisador pode utilizar métodos padrões para calcular as estimativas para o modelo proposto. Na mesma linha de raciocínio, Schafer e Graham (2002) afirmam que a principal virtude do método é a simplicidade, pois o problema da falta de dados pode ser resolvido excluindo-se apenas uma pequena parte da amostra. Dessa forma, pode-se afirmar que o método é completamente eficaz. Entretanto, os autores ressaltam que, mesmo nesta situação, devem-se explorar os dados para se certificar de que os casos descartados não são excessivamente influentes.

Para dados classificados como MCAR, este método é dito como adequado, pois se houver um número de amostras suficiente e o padrão dos valores faltantes for completamente aleatório, então a solução *listwise* pode ser considerada adequada. Da mesma forma, segundo Pigott (2001), se os dados são MCAR e há poucas informações faltantes, há uma maior chance de que os dados completos representem a população, produzindo assim resultados coerentes. Para Van Buuren (2012), se os dados são MCAR, a exclusão *listwise* produz estimativas não tendenciosas para as médias, variâncias e pesos da regressão.

No entanto, a rejeição de casos incompletos pode ser um desperdício desnecessário. Na prática, não é incomum que boa parte da amostra real seja perdida, principalmente se há um grande número de variáveis no estudo. Ou também, de acordo com Mcknight at al. (2007), é bem possível que grande parte da amostra em estudo esteja com muitos dados faltantes em uma mesma variável, e com a eliminação dos casos incompletos, perde-se toda ou grande parte das informações sobre a variável em questão.

Ainda sobre a perda de amostra, King et al. (2001) afirmam que, apesar de, em média, um pouco menos de um terço das observações serem perdidas quando o método *listwise* é aplicado, a proporção de observações perdidas pode ser muito maior. Em seus estudos apresentados na reunião anual da Sociedade para Metodologia Política, em 1997, por exemplo, o número de casos perdidos ultrapassou cerca de 50% em média, e, em alguns casos, mais de 90%.

Além disso, se os dados não são MCAR, os resultados podem ser ainda mais distorcidos ao se aplicar esse método de exclusão. Conforme Acock (2005), se os dados não atendem a suposição de serem MCAR, a exclusão *listwise* pode gerar estimativas tendenciosas, pois geralmente os casos completos podem não ser representativos para

a população. Litlle e Rubin (2002, apud VAN BUUREN,2012) argumentaram que é difícil formular "regras de ouro" para o uso ou não deste método, pois as consequências de seu uso não dependem somente da proporção de dados faltantes.

3.3.2 Método de Exclusão *Pairwise*

O método de exclusão pairwise, também conhecido como Análise dos Casos Disponíveis, ao contrário do método listwise, utiliza todos os casos disponíveis nas bases de dados para estimar os parâmetros do modelo. Assim, o método em questão tenta corrigir o problema de perda de informações da técnica *listwise*, ao eliminar todos os casos com informação faltante. Van Buuren (2012), descreve como o método *pairwise* funciona da seguinte forma:

> O método calcula as médias e as (co) variâncias em todos os dados observados. Assim, a média da variável X é baseada em todos os casos com os dados observados em X, a média da variável Y utiliza todos os casos com os valores de Y observados, e assim por diante. Para as correlações e covariâncias, todos os dados são tomados em que X e Y têm valores não faltantes. Posteriormente, a matriz de estatísticas sumárias é alimentada em um programa de análise de regressão, análise fatorial ou outros procedimentos de modelagem. (VAN BUUREN, 2012, p. 9).

De acordo com Acock (2005), neste método utilizam-se todas as informações possíveis no sentido de que todos os participantes que responderam a um par de variáveis são usados para estimar a covariância entre essas variáveis, independente de eles responderem outras variáveis ou não. Sendo assim, o método é considerado simples e alguns softwares, tais como SPSS, SAS e Stata, contêm diversos procedimentos com opção de se usar o método *pairwise*.

Se os dados são MCAR, o método produz estimativas consistentes das médias, variâncias e covariâncias. O desempenho deste método na análise de casos completos em relação à análise de casos disponíveis, com dados MCAR, irá depender da relação existente entre as variáveis. No método de casos disponíveis, por exemplo,

só serão produzidas estimativas consistentes quando as variáveis forem fracamente correlacionas.

Um dos motivos que faz com que essa técnica não seja muito utilizada na prática é a possibilidade de ela produzir matrizes de covariâncias não plausíveis. Isso se dá, especificamente, pelo fato de que cada covariância pode se basear em subamostras de diferentes participantes. Assim, os erros nas estimativas podem ser causados por causa da diferença entre os números de informações utilizados para calcular os componentes das matrizes de covariâncias. De acordo Schafer e Graham (2002), uma outra limitação do método *pairwise* é que, pelo fato de os parâmetros serem estimados a partir de diferentes conjuntos de observações, é difícil calcular os erros padrão ou outras medidas de incertezas.

Além disso, para dados não MCAR, as estimativas podem ser tendenciosas e nesta situação podem existir ainda problemas computacionais. Van Buuren (2012) afirma ainda que a matriz de correlação pode não ser positiva definida, sendo este um requisito da maioria dos procedimentos multivariados. Assim, a grande dificuldade em usar a análise de casos disponíveis está no fato de que não se pode prever quando este método de exclusão fornecerá resultados adequados ou não, tornando-o assim, não muito útil como um método geral.

3.4 Contextualização dos Dados Faltantes neste estudo

Neste estudo, será feita uma análise de um banco de dados incompleto, o qual iremos supor ter sido gerado a partir de uma aplicação de uma prova para a avaliação de determinada(s) habilidade(s), por exemplo, uma prova de vestibular do CEBRASPE (Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos). Considera-se que, por tratar-se de uma prova, o banco de dados oriundo deste tipo de situação, na maioria das vezes, é composto por dados faltantes. Esses dados faltantes, por sua vez, podem ser causados por vários motivos: falta de tempo para terminar a prova, falta de conhecimento da questão, falta de interesse do aluno em terminar a prova toda, prova cansativa, entre outros.

Tendo em vista os motivos que possam gerar os dados faltantes, neste estudo trabalharemos com a suposição de dados faltantes classificados como MNAR (vide seção 3.2.3), visto que estes dados faltantes dependem de eventos que o pesquisador não consegue observar e controlar. Além desses motivos que possam ter gerado o banco de dados incompleto, consideraremos, ainda, uma outra possibilidade de justificativa da não resposta, já explicada no capítulo anterior e que será retomada a seguir.

Neste trabalho não será realizada a imputação de dados faltantes, visto que, como já foi dito, a informação faltante será tratada de forma diferente: suporemos que *a falta de resposta à questão pelos alunos como uma opção de resposta*, visto que nas provas realizadas pelo CEBRASPE, quando o aluno não sabe a questão, há vantagens em deixá-la em branco, ao invés de "chutar" uma resposta qualquer e correr o risco de errar. Essa vantagem surge conforme a regra de correção das provas do CEBRASPE, em que cada questão incorreta anula uma questão correta. Assim, a não resposta neste banco de dados não será considerada, "necessariamente", um dado faltante, mas sim uma opção de resposta, de acordo com a suposição feita neste trabalho.

Capítulo 4

Metodologia

4.1 Cenário de Estudo

Em banco de dados relacionados à área educacional, mais especificamente, em banco de dados em que são armazenados respostas de provas ou testes que meçam algum tipo de habilidade, às quais tenham a característica de não haver penalidade extra às questões incorretas, como por exemplo, a prática de anular-se uma questão correta para cada questão incorreta, a tendência é que se tenha um banco de dados completo. Em outras palavras, pode-se afirmar que, neste caso, é incomum a presença de dados faltantes. Mesmo na situação em que, por exemplo, falte tempo para o respondente resolver todas as questões, há a possibilidade de "chutar" essas questões, obtendo assim, respostas para todas as questões dos indivíduos.

Entretanto, quando na prova há o referido sistema de penalização caso o indivíduo erre a questão, como a perda de pontos na nota final, há uma tendência à ocorrência de dados faltantes. Tal premissa é baseada na suposição de que o respondente saiba que há vantagem em não responder a questão, caso ele não tenha certeza da resposta.

Um exemplo de prova com esta característica de penalidade para respostas incorretas é a prova do vestibular da Universidade de Brasília (UnB), a qual é elaborada pelo Centro Brasileiro de Pesuisa e Seleção e de Promoção de Eventos (CESBRASPE).

As provas objetivas deste teste, como mencionado no capítulo 1, são caracterizadas por quatro tipos de questões, denominados tipo A, tipo B, tipo C e tipo

D. De forma resumida, as questões são definidas da seguinte forma: as questões do tipo A tem como opções de respostas "Certo" ou "Errado"; nas questões do tipo B, é proposto um problema ao candidato e o mesmo deve marcar um único resultado numérico como resposta da questão, representado por um número inteiro de 000 a 999; as questões do tipo C têm quatro opções de respostas, designadas pelas letras A, B, C e D, das quais apenas uma constitui a resposta correta; e por fim, as questões do tipo D são itens de respostas construídas, ou seja, questões abertas, com respostas elaboradas pelo candidato.

Neste estudo, como também já citado anteriormente, iremos considerar que o banco de dados será constituído com questões do tipo A em que, além destas características citadas anteriormente, elas são corrigidas a partir do seguinte cálculo: Caso a resposta do candidato esteja em concordância com o gabarito oficial definido na prova, ou seja, ele acerte a questão, ele tem uma pontuação +1 (um ponto positivo). Caso a resposta do candidato esteja em discordância com o gabarito oficial definido na prova, ou seja, ele erre a questão, ele tem uma pontuação -1 (um ponto negativo). E, por fim, caso não haja marcação por parte do candidato, ele tem pontuação 0 (zero pontos).

Assim, é constatado que, a cada resposta errada que o candidato marcar, ele irá anular uma resposta certa que ele já tenha respondido ou venha a responder. Percebe-se que, em casos de incertezas do candidato quanto à resposta correta da questão, a não resposta costuma ser vantajosa. Sendo assim, no banco de dados utilizado nesta pesquisa, haverá três opções de respostas: "Errar", "Não Responder" e "Acertar".

Para a obtermos um banco de dados com as características citadas anteriormente, foi simulado um conjunto de informações através de alguns passos, os quais serão descritos na próxima seção. Além de analisarmos o banco de dados simulados, iremos também analisar um banco de dados real do vestibular da UnB, o qual será descrito na seção 4.3 deste capítulo.

4.2 Simulação do Banco de Dados

Para que o banco de dados utilizado neste estudo tenha as características descritas anteriormente, foram realizados os seguintes passos em sua construção:

1º Passo:

Foram gerados 50 valores aleatoriamente a partir da distribuição Uniforme no intervalo [0.5,3], referente aos valores dos parâmetros de discriminação dos itens (a_i). Em seguida, foram gerados também, 50 valores distribuídos de forma equidistantes no intervalo [-1,5,1,5], referente aos valores dos parâmetros de dificuldade dos itens (b_i). Após isto, foram gerados n=1000 valores referentes às proficiências de 1000 indivíduos θ_j através de uma distribuição Normal (0,1).

2º Passo:

Foram gerados n=1000 valores de uma variável M a partir de uma distribuição Beta (1,2;0,8), gerada a partir de uma média 0,8 e de um desvio padrão igual a 0,20, a qual representa um ponto de corte em que determina se o indivíduo vai ou não, primeiramente, responder a questão. Assim, será determinado um valor M_j, para cada respondente j, em que j varia de 1 a 1000. O valor de M_j indica quão propenso ao "chute" é o candidato. Quanto maior o valor de M_j, menos propenso ao "chute" o candidato é.

Os parâmetros a e b da distribuição Beta são determinados de tal forma que, se Y $\sim$ Beta (a, b), então

$$E(Y) = \frac{a}{a+b} = 0,8 \qquad e \qquad Var(Y) = \frac{ab}{(a+b)^2(a+b+1)} = 0,04$$

3º Passo:

A partir dos valores dos parâmetros dos itens gerados no primeiro passo e da variável M gerada no segundo passo, foi utilizado o modelo da TRI de dois parâmetros para gerar respostas binárias (0,1) para uma variável aleatória V_{ij}, com $i = 1, ...50$

e $j = 1, ..., 1000$, que indica se o *j-ésimo* indivíduo vai responder a *i-ésima* questão. Sendo assim, $P(V_{ij} = 1|\theta_j)$ é a probabilidade do *j-ésimo* indivíduo responder a *i-ésima* questão e esta probabilidade será dada pelo modelo da TRI de dois parâmetros, com θ_j descrevendo a proficiência do indivíduo *j*.

Seja U_{ij} uma variável aleatória que assume o valor 1 se o *j-ésimo* indivíduo acerta a *i-ésima* questão, 0 se o *j-ésimo* indivíduo não responde a *i-ésima* questão e -1 se o *j-ésimo* indivíduo erra a *i-ésima* questão.

Portanto, se $P(V_{ij} = 1|\theta_j) \geq M_j$, foi gerada uma variável aleatória X seguindo uma distribuição Uniforme (0,1) e foi definido U_{ij}= 1 se $X \leq P(V_{ij} = 1|\theta_j)$ ou U_{ij}= -1 se $X > P(V_{ij} = 1|\theta_j)$. Caso contrário, se $P(V_{ij} = 1|\theta_j) < M_j$, foi definido então que U_{ij}=0.

Ou seja, dado que foi definido no passo 3 se ele iria responder a questão ou não, caso ele respondesse, ele iria ter ainda a possibilidade de errar ou acertar. Por outro lado, caso ele não respondesse, por não ter certeza da resposta da questão, ou seja, por não ter "habilidade" suficiente sobre o traço latente medido, ele iria gerar uma "não resposta" ou seja, um dado faltante.

Após estes passos, de forma resumida, obtemos um banco de dados com valores -1, indicando que o individuo errou a questão; 0, indicando que o individuo não respondeu a questão e; 1, indicando que o individuo acertou a questão. A Figura 4.1 ilustra através de uma "Árvore de probabilidade" os passos descritos acima para a simulação do banco de dados.

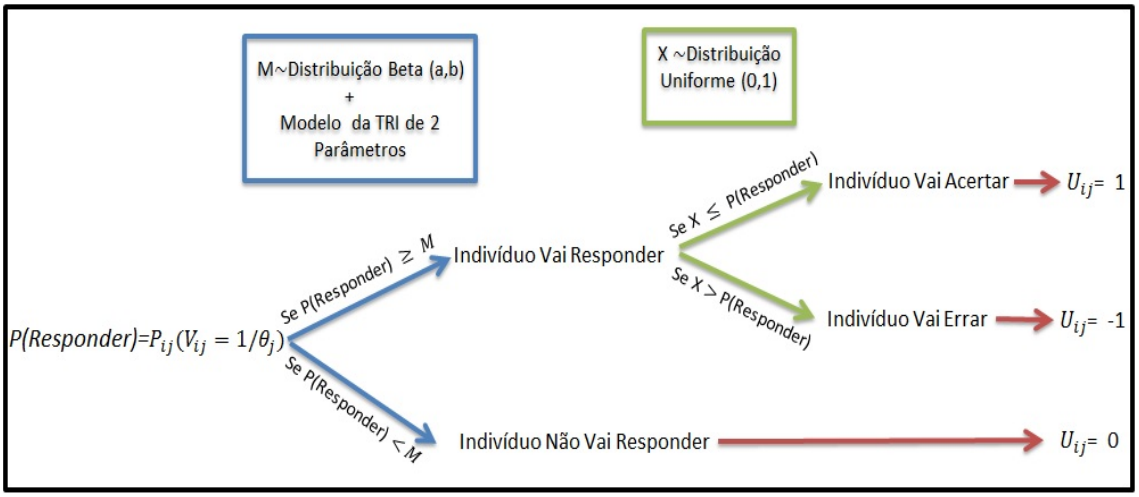

Figura 4.1: Simulação dos Dados. Fonte: Elaborado pela autora.

A título de exemplificação, na Figura 4.2 estão ilustradas as curvas características de dois itens hipotéticos (Item 1 e Item 2) e a proficiência θ_j de um indivíduo j. Supondo que o valor da variável M gerada para este indivíduo tenha sido 0,8, temos que:

- Para o item 1, a probabilidade de o indivíduo responder a questão foi igual a 0,88, ou seja, $P(V_{1j} = 1|\theta_j)$=0,88. Como o valor da probabilidade de o indivíduo responder o item 1 é maior que $M_j = 0,8$, então o indivíduo vai responder a questão, podendo acertar ou errar. Neste caso, gera-se a variável X seguindo uma uniforme (0,1) e se X for maior que 0,88, será gerado no banco de dados o valor -1, indicando que o indivíduo errou a questão. Caso contrário, se X for menor que 0,88, será gerado no banco de dados o valor 1, indicando que o indivíduo acertou a questão.

- Para o item 2, o valor da probabilidade de o indivíduo responder a questão resultou em 0,27, ou seja, $P(V_{2j} = 1|\theta_j)$=0,27. Como o valor da probabilidade de responder dele é menor que $M_j = 0,8$, constata-se que o indivíduo não vai responder a questão, gerando assim, o valor 0 no banco de dados.

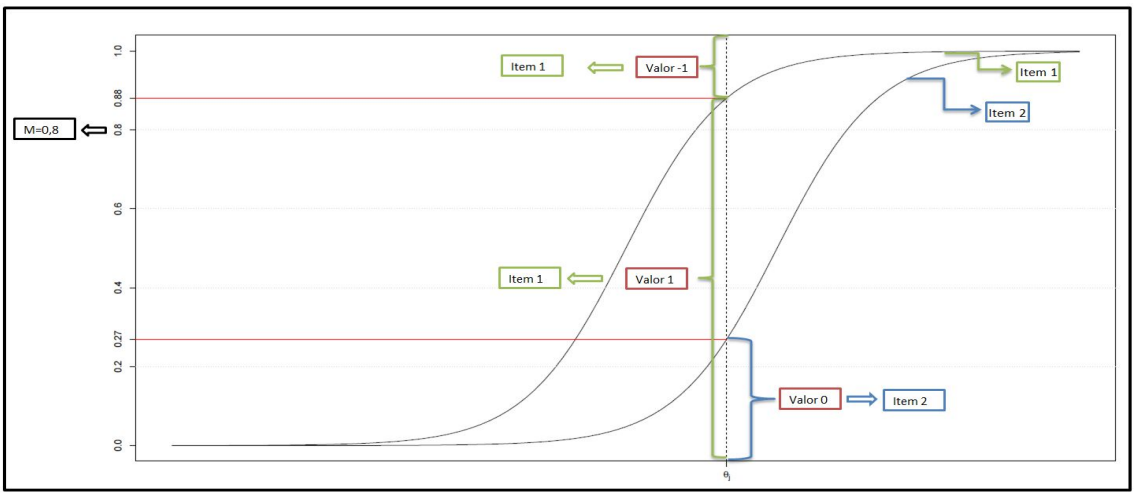

Figura 4.2: Gráfico de um exemplo hipotético da CCI de dois itens distintos.
Fonte: Elaborado pela autora.

Além desse banco de dados, foram simuladas também outras situações, variando o número de indivíduos em 5.000, 10.000 e 20.000 candidatos, e o número de itens, em 50, 100, 150 e 200 itens. Estes casos foram gerados com o intuito de avaliarmos diferentes situações para a comparação dos ranks dos candidatos obtidos através do método convencional e dos ranks obtidos pelo modelo de resposta gradual. Esse assunto será melhor detalhado na Seção 5.4 do capítulo a seguir.

4.3 Dados Reais

Com o intuito de aplicarmos também a metodologia proposta considerando uma situação real, além do estudo de simulação, analisamos um banco de dados real disponibilizado pelo CESBRASPE.

Esse banco de dados continha provas dos vestibulares da UnB referente aos anos/semestre de 2013/2, 2014/2, 2015/2. No banco de dados de 2013/2, havia 21.311 indivíduos, no de 2014/2 havia 21.968 indivíduos e no de 2015/2 havia 17.242 indivíduos. Em cada um dos três anos, os participantes responderam 300 questões ao todo. Neste trabalho foram analisadas somente as informações referentes ao vestibular de 2014/2. A seguir, apresentamos um quadro descritivo do vestibular de 2014/2.

Data	Prova	Disciplinas-Foco	Nº de itens	Duração
1º DIA 7/6/2014	Conhecimentos – Parte I	Língua Espanhola, Língua Francesa ou Língua Inglesa	30	300min
	Conhecimentos – Parte II	Língua Portuguesa e Literaturas de Língua Portuguesa, Geografia e História, Artes (Artes Cênicas, Artes Visuais e Música), Filosofia e Sociologia	120	
	Redação em Língua Portuguesa	–	–	
2º DIA 8/6/2014	Conhecimentos – Parte III	Biologia, Física, Química e Matemática	150	300min

Figura 4.3: Características da prova do vestibular 2014/2.
Fonte: Edital Nº 1 do Vestibular da UnB de 2014 (CEBRASPE)

Conforme mostrado na Figura 4.3, o vestibular de 2014/2 foi realizado em dois dias (7 e 8 de junho) e a prova aplicada foi dividida em três partes. No primeiro dia, foi aplicada a 1ª parte da prova, correspondente às questões referentes à Língua Estrangeira (Inglês, Espanhol e Francês), totalizando 30 itens, e à 2ª parte, a qual se caracterizou por conter as questões de Português e Literaturas (50 questões), Geografia (24 questões), História (30 questões), Artes (8 questões), Filosofia (3 questões) e Sociologia (5 questões). Sendo assim, no primeiro dia, os participantes responderam 150 questões em 5 horas, além de elaborarem uma redação.

No segundo dia, foi aplicada a 3ª parte da prova, a qual corresponde a aplicação das questões referentes às áreas de Biologia (38 questões), Física (36 questões), Química (39 questões) e de Matemática (37 questões), totalizando também 150 questões aplicadas.

Analisando os dados disponibilizados pelo CEBRASPE, percebemos que havia três tipos de provas diferentes. Cada tipo continha as mesmas questões e quantidades, entretanto se diferenciava pela ordem em que as questões foram dispostas, não havendo nenhum padrão de ordenação, nem de forma geral e nem dentro de cada assunto/área abordada. Em outras palavras, a primeira questão da prova do tipo I, não correspondia à primeira questão da prova do tipo II e/ou III, nem mesmo com relação à área cobrada, caso a ordem mudasse somente dentro de cada assunto dis-

cutido. Assim, dentro das informações do vestibular de 2014/2, foi identificado que 7.347 indivíduos responderam a prova do tipo I, 7.307 responderam a prova do tipo II e 7.312 responderam a prova do tipo III.

Além disso, em cada tipo de prova (tipo I, tipo II, tipo III), por sua vez, na parte 1, parte em que foi abordada a língua estrangeira, havia mais três opções de provas referentes às provas de Inglês, Francês e Espanhol. Dessa forma, dentro de cada tipo de prova, havia indivíduos que tinham respondido a prova de inglês, outros de espanhol e outros de francês. Sendo assim, ao invés de três tipos de provas diferentes, havia, na verdade, 9 tipos de provas distintas.

Depois dessa constatação, para que fosse possível realizar a análise de forma correta, as questões referentes às provas de língua estrangeira foram desconsideradas (30 questões). Foram dispensadas também as provas de tipo II e III. Assim, foram analisados somente os indivíduos que fizeram a prova do tipo I e somente as questões da 2ª e 3ª partes da prova. Dessa forma, conseguimos obter um exemplar de prova homogêneo em que todos os indivíduos responderam exatamente as mesmas questões.

Dentre os indivíduos que responderam a prova do tipo I, no entanto, foi observado que houve 115 indivíduos que fizeram a prova somente no primeiro dia, o que fez com que o banco de dados ficasse incompleto com relação às respostas das questões do segundo dia. Foi observado também que para dois indivíduos o gabarito da prova do primeiro dia era referente à prova do tipo I e do segundo dia era da prova do tipo II. Dessa forma, não foi possível identificar qual tipo de prova estes indivíduos responderam. Assim, os indivíduos presente nesses casos foram desconsiderados da análise.

Como apresentado na seção 4.1 deste trabalho, as provas do vestibular do CEBRASPE contêm 4 tipos de questões: A, B, C e D. No vestibular de 2014/2, para o primeiro dia de prova, havia 113 questões do tipo A, 0 questões do tipo B, 6 questões do tipo C e 1 questão do tipo D, já desconsiderando as questões referentes à língua estrangeira. Já para o segundo dia de prova, haviam 132 questões do tipo A, 4 questões do tipo B, 13 questões do tipo C e 1 questão do tipo D. Assim, somando as questões do tipo A, a qual é o tipo de questão proposta a ser analisada neste trabalho, do primeiro e segundo dia, estas totalizaram 245 questões. Entretanto, 3 questões do primeiro dia de prova e 5 questões do segundo dia foram anuladas. Dessa forma,

foram utilizadas na análise um total de 237 questões.

Em síntese, foram analisadas as 1.713.984 respostas de 7.232 indivíduos às 237 questões do vestibular da UnB de 2014/2 elaborado pelo Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (CEBRASPE). Os resultados da análise desse banco de dados real serão apresentados posteriormente no capítulo 6.

Capítulo 5

Resultados - Dados Simulados

Neste capítulo, iremos analisar os resultados gerados a partir da simulação dos dados descrita na seção 4.2. O programa utilizado para gerar os resultados foi o Software R e a função utilizada para aplicar o modelo de resposta gradual (MRG) nos dados foi a função *mirt* do pacote **mirt**. Para o cálculo dos escores dos indivíduos, sendo estes as notas dos indivíduos estimadas pelo MRG, foi utilizada a função *fscores* e para a estimação dos parâmetros dos itens foi utilizado a função *coef*.

5.1 Análise Descritiva

Conforme apresentado nos passos 1, 2 e 3 da seção 4.2, foram gerados valores dos parâmetros de dificuldade e discriminação com o intuito de aplicar o modelo da TRI de dois parâmetros, gerando assim respostas binárias (0,1) para a variável aleatória V_{ij}, a qual indica se o j-ésimo indivíduo irá, primeiramente, responder a i-ésima questão ou não. Em seguida, a probabilidade de o indivíduo responder a i-ésima questão dado a sua habilidade, $P(V_{ij} = 1|\theta_j)$, foi comparada com os valores gerados para a variável M_j, gerando assim, o banco de dados final, com $U_{ij} = 1$, indicando que o indivíduo acertou a questão, com $U_{ij} = 0$, se o indivíduo não respondeu a questão e, por fim, com $U_{ij} = -1$, se o indivíduo errou a questão.

Através da Tabela 5.1, é possível observar que, para os valores referentes ao parâmetro de discriminação, o valor mínimo resultou em 0,5861, indicando que houve itens com baixa discriminação, baseando-se no ponto de corte que a literatura

indica, sendo este maior que 1. Por outro lado, podemos observar também que houve itens que descriminaram bem os indivíduos, visto que o valor máximo resultou em 2,96. Apesar de termos itens com baixa discriminação, em média obtivemos itens com índices de discriminação em torno de 1,73, indicando que, em média, o teste como um todo fará uma boa diferenciação entre os indivíduos baseado na escolha de responder ou não a questão.

Tabela 5.1: *Análise Descritiva para os parâmetros dos itens simulados, Proficiências e M_j (Dados Simulados)*

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Parâmetros de Discriminação	0,5861	1,2340	1,7960	1,7380	2,2780	2,9600
Parâmetros de Dificuldade	-1,5000	-0,7500	0,0000	0,0000	0,7500	1,5000
Proficiências	-3,7800	-0,6882	-0,0681	-0,0393	0,5969	2,9510
M_j, j=1,2,...,1000	0,0034	0,3697	0,6200	0,5893	0,8312	0,9998

Em relação aos parâmetros de dificuldade dos itens, observou-se um mínimo de -1,5 e um máximo de 1,5, o que era de se esperar pelo método de geração desses valores (valores equidistantes no intervalo de -1,5 a 1,5). Isso indica também que terão itens mais difíceis ou menos difíceis, indicando uma boa avaliação dos indivíduos que escolherão responder ou não.

Quanto aos valores gerados para as proficiências dos indivíduos, observou-se um mínimo de -3,78, indicando que há indivíduos com habilidades baixas, o que de fato acontece na maioria das provas que objetivam medir algum tipo de conhecimento, pois muitos indivíduos não respondem a questão em razão de não terem alto conhecimento do assunto abordado. De outro modo, observaram-se também indivíduos com alta habilidade (máximo igual a 2,95), indicando, da mesma forma, casos que ocorrem com frequência nesse tipo de teste, pois sempre há, também, indivíduos que irão, com certeza, responder a determinada questão por terem um alto conhecimento no assunto.

Para os valores da variável M, a qual se refere ao ponto de corte em que determina se o indivíduo vai ou não responder a questão, obtivemos um mínimo de 0,0089 e um máximo de 0,9999. Esses valores foram gerados, como já mencionado, de tal forma que a distribuição de seus valores seguissem uma Beta (1,2; 0,8), baseado numa média igual a 0,8 e um desvio padrão igual a 0,20. Esses valores foram escolhidos

com o objetivo de se ter tantos valores muito baixos de M, como valores muito altos de M.

Baseado nesses valores, conseguimos colocar diferentes situações que imaginemos que aconteçam em uma situação real. Por exemplo, obtendo valores da variável M muito altos, tanto para indivíduos com alta habilidade quanto para indivíduos com baixa habilidade, estaremos cobrindo tanto indivíduos que irão responder a questão por realmente saberem do assunto e serem cautelosos, representando os indivíduos que não "chutam" nada, como também indivíduos que não irão responder a questão por não saberem a resposta correta e também por serem muito cautelosos. Por outro lado, obtendo valores de M muito baixos, tanto para indivíduos de alta habilidade como para indivíduos de baixa habilidade, estaremos cobrindo tanto indivíduos que irão responder a questão por realmente saberem do assunto, porém que não são cautelosos, correndo o risco de errar por responderem uma questão sem ter tanta certeza da resposta correta, quanto indivíduos que irão responder a questão por não saberem do assunto e também não serem cautelosos, retratando, por exemplo, aqueles indivíduos que "chutam" tudo.

A título de exemplificação, na Tabela 5.2 estão apresentadas as frequências das respostas de alguns indivíduos. Podemos observar, por exemplo, que o indivíduo 6 pode representar um respondente regular, com um conhecimento médio sobre o assunto e que também deva possuir um valor de M médio, nem alto, nem baixo. Já o individuo 7 pode representar aquele aluno que não sabe muito do assunto abordado e é um pouco cauteloso, pois das 27 questões que respondeu, ele errou 22, mas ainda assim, deixou de responder 23 questões.

Tabela 5.2: *Frequências de Respostas por Indivíduo (Dados Simulados).*

	Respostas		
Indivíduos	-1	0	1
1	15	23	12
2	1	34	15
6	13	23	14
7	22	23	5
8	9	28	13
9	3	32	15
11	-	41	9
13	4	27	19
15	5	31	14
16	15	23	12
18	-	50	-
19	2	45	3
56	7	-	43
677	36	1	13
711	-	15	35

Os indivíduos 19 e 11 representam alunos que não têm muito conhecimento sobre o assunto, porém são muito cautelosos, não "chutam" muitas questões e só respondem quando têm muita certeza. Podemos concluir isso pelo fato de não terem acertado quase nada. Entretanto, não erram praticamente nada (Individuo 19) ou realmente nenhuma questão (Indivíduo 11). Para o indivíduo 15, as quantidades estão de acordo com um aluno que sabe um pouco do assunto e é cauteloso, pois este errou somente 5 questões, deixando 31 em branco, e acertou 14 questões.

Para o indivíduo 18, podemos perceber que ele pode ser um aluno que não sabe nada, mas também não "chuta" nada, não fazendo nenhum item da prova. Já o indivíduo 56, representa um aluno que tem um alto conhecimento do assunto, acertando muitas questões (43 itens), porém não é cauteloso, pois não deixou de responder questão alguma e errou 7 questões, que talvez não tivesse tanta certeza da resposta e quis "chutar", independente disso.

O indivíduo 677, aparentemente, é um aluno que não tem muito domínio do conteúdo avaliado e, além disso, não é muito cauteloso, pois respondeu praticamente todas (49 questões) e errou mais da metade das questões (36). Já o indivíduo 711 representa um aluno com um alto conhecimento do assunto medido na prova, acertando um número de questões razoável (35) e é extremamente cuidadoso, pois não

errou nenhuma questão, deixando 15 questões em branco.

Na Tabela 5.3, está apresentada a frequência de respostas global do banco de dados simulado após a comparação da variável M_j de cada indivíduo com sua respectiva probabilidade de acertar determinada questão, oriunda do modelo da TRI de dois parâmetros, conforme já descrito. Num total de 50.000 respostas, correspondente às respostas de 1000 indivíduos a 50 questões de uma prova, observamos que aproximadamente 10% das questões foram marcadas de forma errada, aproximadamente 60% das questões não foram respondidas e aproximadamente 30% das questões foram respondidas corretamente.

Tabela 5.3: *Frequências de Respostas (Dados Simulados).*

	Respostas		
	-1	0	1
Frequência	5260	29892	14848
Percentual	10,52%	59,78%	29,69%

Consideramos esses valores razoáveis, visto que a aplicação de uma penalização caso o indivíduo erre a questão, faz com que grande parte dos indivíduos não responda questões para as quais não tenham tanta certeza da resposta correta ou que realmente não saibam. Dessa forma, um número elevado de valores de zeros no banco de dados era esperado. Da mesma forma, valores baixos de respostas corretas (1) no banco de dados eram esperados, mas não tanto quanto os valores de respostas erradas (-1), pois, como já citado, os indivíduos tinham consciência de que não responder poderia ser mais vantajoso, evitando a soma de um valor negativo em sua nota final. Essas frequências estão representadas graficamente na Figura 5.1.

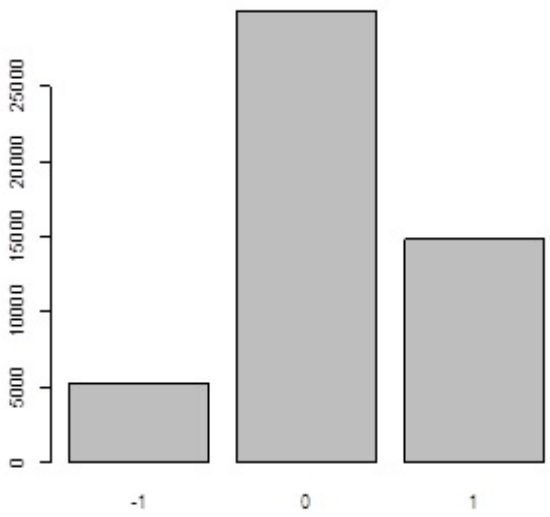

Figura 5.1: Histograma das frequências de respostas (Dados Simulados).

5.2 Análise dos Parâmetros dos Itens e Curva Característica dos Itens

Na Tabela 5.4, estão apresentados os valores dos parâmetros dos itens estimados e a frequência de respostas dos indivíduos por item. A partir desses resultados, podemos observar que, por exemplo, a quantidade de acertos do item 1 resultou em 51%, 35,3% dos indivíduos não responderam essa questão e apenas 13,7% erraram. Analisando o índice de discriminação deste item, pode-se concluir que ele discrimina de forma razoável os indivíduos e, ao se observar os valores dos parâmetros de dificuldade, observamos valores negativos de suas categorias, resultando ainda um valor negativo da média entre elas. Podemos concluir então que este seria um item relativamente fácil.

Observando agora o item 3, observamos que foi o item que os indivíduos mais acertaram. Apesar de haver indícios de boa discriminação dos indivíduos (a_3=6,526), percebe-se que ele é um item muito fácil, pois tem baixos valores dos parâmetros de dificuldade, fazendo com que muitos indivíduos acertem a questão (72,9% de acerto).

Tabela 5.4: *Frequência das respostas por Item e Parâmetros dos itens (Dados Simulados).*

Itens	Categoria -1 (%)	Categoria 0 (%)	Categoria 1 (%)	a_i	b_{i1}	b_{i2}	$b_{\text{i,médio}}$
Item 1	13,700	35,300	51,000	1,168	-1,941	-0,232	-1,2025
Item 2	14,300	32,600	53,100	1,971	-1,349	-0,324	-0,9985
Item 3	7,300	19,800	72,900	6,523	-1,043	-0,689	-1,2105
Item 4	17,300	43,800	38,900	1,076	-1,793	0,314	-0,5825
Item 5	7,900	25,600	66,500	3,954	-1,193	-0,608	-1,2045
Item 6	12,900	33,600	53,500	2,239	-1,323	-0,341	-1,0025
Item 7	10,500	25,500	64,000	4,038	-1,108	-0,571	-1,1250
Item 8	9,500	25,000	65,500	5,157	-1,053	-0,592	-1,1185
Item 9	11,800	34,000	54,200	2,743	-1,239	-0,370	-0,9895
Item 10	10,700	27,900	61,400	4,294	-1,078	-0,528	-1,0670
Item 11	15,300	43,600	41,100	1,556	-1,499	0,085	-0,6645
Item 12	12,200	39,600	48,200	1,862	-1,507	-0,188	-0,9415
Item 13	17,700	49,200	33,100	0,855	-2,108	0,753	-0,3010
Item 14	17,400	47,400	35,200	1,033	-1,847	0,499	-0,4245
Item 15	12,600	42,400	45,000	2,399	-1,292	-0,139	-0,7850
Item 16	13,800	45,700	40,500	1,625	-1,546	0,084	-0,6890
Item 17	11,900	43,700	44,400	2,529	-1,293	-0,136	-0,7825
Item 18	14,000	46,900	39,100	1,904	-1,401	0,079	-0,6215
Item 19	10,700	47,100	42,200	2,736	-1,290	-0,090	-0,7350
Item 20	15,500	53,500	31,000	1,219	-1,751	0,615	-0,2605
Item 21	9,200	48,700	42,100	3,003	-1,303	-0,104	-0,7555
Item 22	12,000	54,100	33,900	2,107	-1,413	0,222	-0,4845
Item 23	11,100	54,100	34,800	2,450	-1,350	0,143	-0,5320
Item 24	13,100	57,400	29,500	1,851	-1,465	0,434	-0,2985
Item 25	13,800	59,600	26,600	1,489	-1,633	0,686	-0,1305
Item 26	17,300	60,400	22,300	0,719	-2,465	1,740	0,5075
Item 27	9,400	62,200	28,400	1,978	-1,610	0,456	-0,3490
Item 28	10,100	64,300	25,600	1,685	-1,733	0,658	-0,2085
Item 29	10,100	66,000	23,900	1,512	-1,861	0,814	-0,1165
Item 30	7,900	67,700	24,400	2,073	-1,657	0,619	-0,2095
Item 31	15,600	66,700	17,700	0,669	-2,803	2,334	0,9325
Item 32	10,700	70,100	19,200	1,251	-2,081	1,255	0,2145
Item 33	7,500	73,300	19,200	1,681	-1,931	1,014	0,0485
Item 34	10,200	73,100	16,700	1,192	-2,203	1,487	0,3855
Item 35	16,900	69,200	13,900	0,329	-5,050	5,525	3,0000
Item 36	6,400	77,900	15,700	1,444	-2,269	1,382	0,2475
Item 37	8,900	78,800	12,300	1,106	-2,464	2,012	0,7800
Item 38	6,500	81,600	11,900	1,319	-2,411	1,826	0,6205
Item 39	14,900	74,000	11,100	0,359	-5,062	5,811	3,2800
Item 40	6,900	83,600	9,500	0,929	-3,136	2,668	1,1000
Item 41	4,500	85,700	9,800	1,414	-2,561	2,003	0,7225
Item 42	6,200	85,900	7,900	0,926	-3,264	2,940	1,3080
Item 43	5,700	86,300	8,000	1,034	-3,059	2,701	1,1715
Item 44	4,100	88,400	7,500	1,126	-3,162	2,627	1,0460
Item 45	6,700	86,400	6,900	0,772	-3,709	3,620	1,7655
Item 46	2,500	91,100	6,400	1,420	-2,980	2,492	1,0020
Item 47	3,300	91,800	4,900	1,075	-3,492	3,243	1,4970
Item 48	3,900	91,300	4,800	0,953	-3,693	3,546	1,6995
Item 49	6,400	89,100	4,500	0,497	-5,639	6,280	3,4605
Item 50	7,200	88,200	4,600	0,175	-14,779	17,318	9,9285

O item 26 pode ser considerado um item relativamente difícil, por ter o seu índice de dificuldade da maior categoria positivo, sendo alto o suficiente para ter também b intermediário entre categorias também maior que zero ($b_{26,\text{médio}}$=0,5075), havendo grande quantidade de não respostas e valores de acertos e erros próximos, porém baixos. Apesar de este item ter sido considerado relativamente difícil, o mesmo não discrimina tão bem os indivíduos (a_{26}=0,719). O item 27 é parecido com o item 26, porém ele discrimina melhor os indivíduos, com parâmetro a_i próximo de 2 (a_{27}=1,978). Apesar de o parâmetro de dificuldade intermediário ser negativo, o valor de b da categoria mais alta é positivo. Dessa forma, temos um item relativamente difícil e com boa discriminação dos indivíduos, resultando em 62,2% de não resposta e 28,4% de acertos.

Os itens 46 e 47 obtiveram pouquíssimas respostas, tanto erradas (2,5% e 3,3%) quanto corretas (6,4% e 4,9%), respectivamente. Provavelmente, estes itens retratem itens contidos em provas, considerados extremamente difíceis ou que abordem um assunto pouco debatido ou usualmente não estudado pelos alunos. Um percentual de não respostas de 91% indica que os indivíduos não quiseram arriscar de forma alguma esse item, pois, baseado no pressuposto já citado, tiveram consciência que teriam poucas chances de acertar o item e não responderam. De fato, os índices de dificuldade intermediários e de discriminação se mostraram realmente de acordo com o resultado da não resposta, sendo estes itens considerados difíceis.

Após analisarmos cada valor da estimação dos parâmetros dos itens, foram gerados seus respectivos gráficos, a Curva Característica do Item (CCI). Conforme já mencionado, a CCI mostra a relação existente entre a probabilidade de o individuo acertar o item dado a sua habilidade, $P(U_{ij} = 1|\theta_j)$, e os parâmetros do modelo.

Para a compreensão dos gráficos, é preciso lembrar que o eixo horizontal representa o valor do traço latente medido (conhecimento sobre determinado assunto) e no eixo vertical corresponde a probabilidade do indivíduo ter sua resposta classificada numa categoria ou superior (para as categorias 0 e 1). Da mesma forma como a Tabela 5.4, é possível ver, agora graficamente, que possuem itens mais discriminantes e/ou mais difíceis que outros.

A Figura 5.2 apresenta a CCI do item 27. Neste gráfico é possível observar que indivíduos com habilidade até aproximadamente -1,6 tem maior probabilidade de errar a questão. Já indivíduos com habilidade entre -1,8 até, aproximadamente, 0,4, tem maior probabilidade de não responder a questão. E por fim, indivíduos com habilidade acima de 1,3, aproximadamente, tem maior probabilidade de acertar a questão. Comparando essas análises com os valores dos parâmetros estimados mostrados na Tabela 5.4, percebe-se que o item de fato tem boa discriminação, possuindo as curvas bem espaçadas entre si, fazendo com que seja possível visualizar a diferença entre respostas dos indivíduos de acordo com a sua respectiva proficiência.

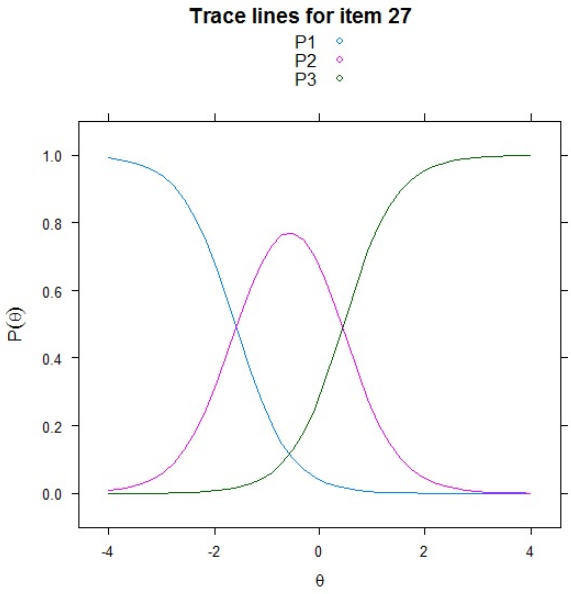

Figura 5.2: Curva Característica do item 27.

Analisando a Figura 5.3, podemos observar que o item 20 é mais difícil que o item 16. Isso ocorre pelo fato de que a habilidade necessária para um indivíduo acertar a questão 20 (aproximadamente 0,6) é maior do que a habilidade necessária para um indivíduo acertar a questão 16 (aproximadamente 0).

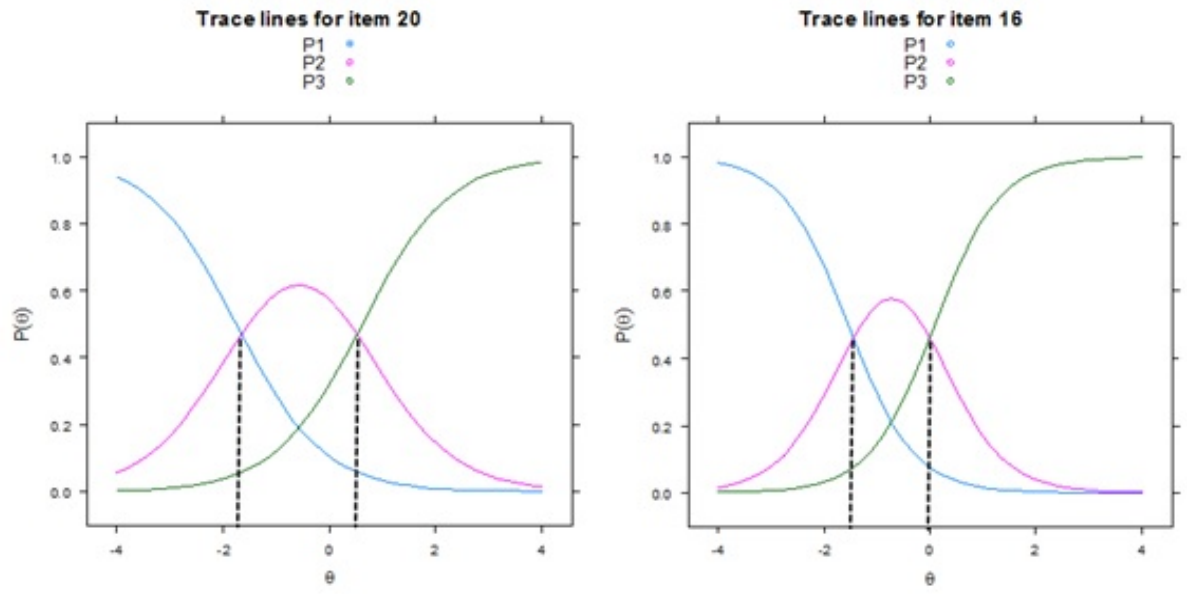

Figura 5.3: Curva Característica dos Itens 20 e 16.

Percebeu-se, também, tanto através das CCIs como através dos valores da Tabela 5.4, que alguns itens não se sobressaíram em nenhuma região do gráfico, indicando que esses itens podem não ter boa qualidade para avaliar o traço latente. Houve também índices dos parâmetros, tanto de discriminação como de dificuldade, com valores muito extremos, se distanciando dos valores ideais. Esses itens poderiam ser representados, por exemplo, pelas questões 3, 39, 49, 50, entre outros. Quando se tem o intuito de elaborar um questionário ou uma prova para depois aplicá-los, a literatura indica que retiremos esses itens não informativos, os quais não avaliam o indivíduo de forma eficiente, e façamos a análise novamente. Entretanto, como este não é o objetivo principal desde trabalho, essa reavaliação não foi realizada. Na Figura 5.4 estão apresentadas as CCIs de todos os itens.

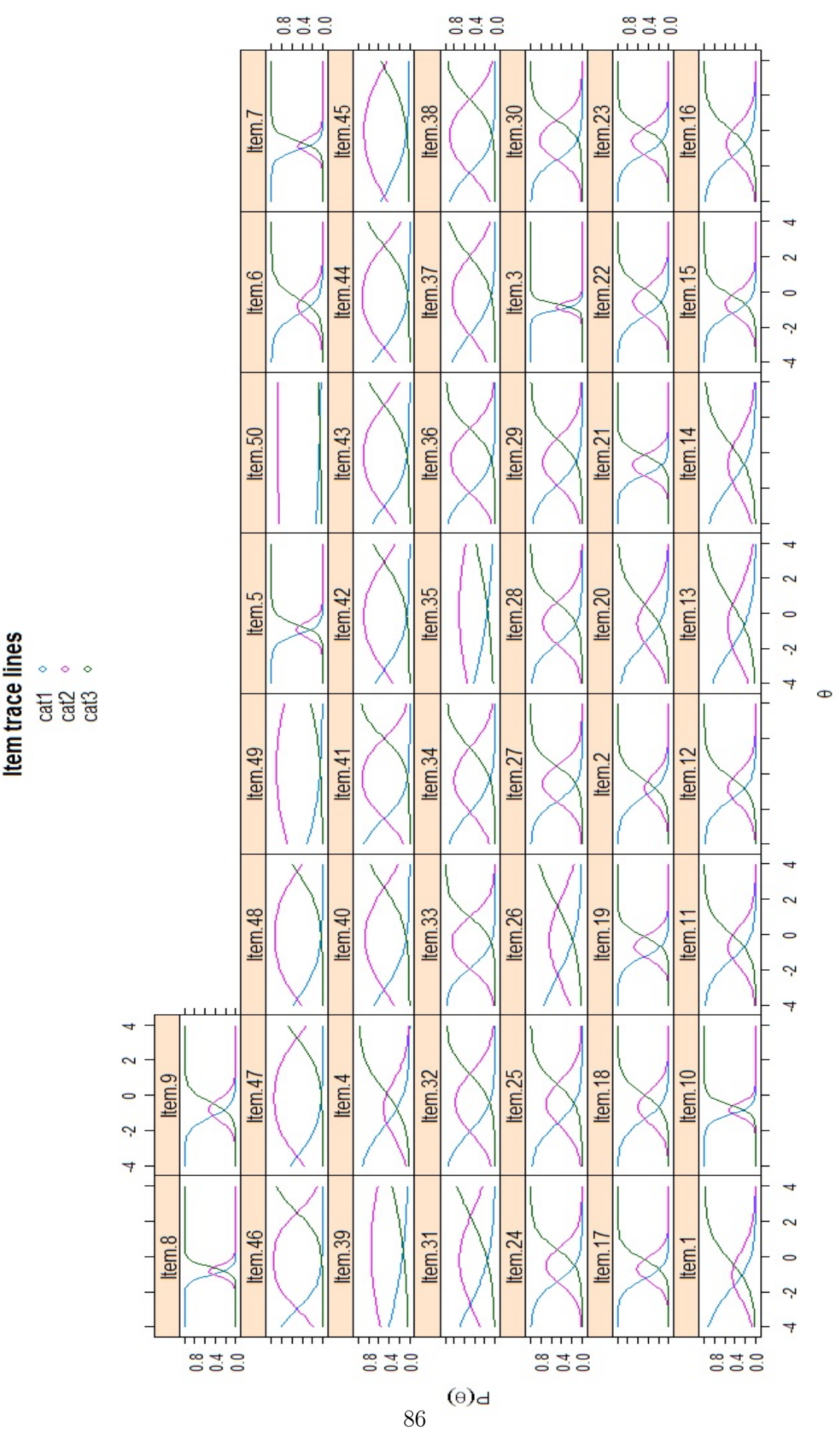

Figura 5.4: Curvas Características dos Itens 1 a 50.

A Figura 5.5 apresenta a Curva de Informação do Teste. Observou-se que o instrumento de medida tem maior informação para os valores da habilidade no intervalo aproximado de -2 à 1. Isso significa, de acordo com Junior et al. (2015), com as devidas adaptações a este trabalho, que o teste como um todo é mais adequado para medir as habilidades de indivíduos que tenham valores das proficiências no intervalo entre -2 à 1.

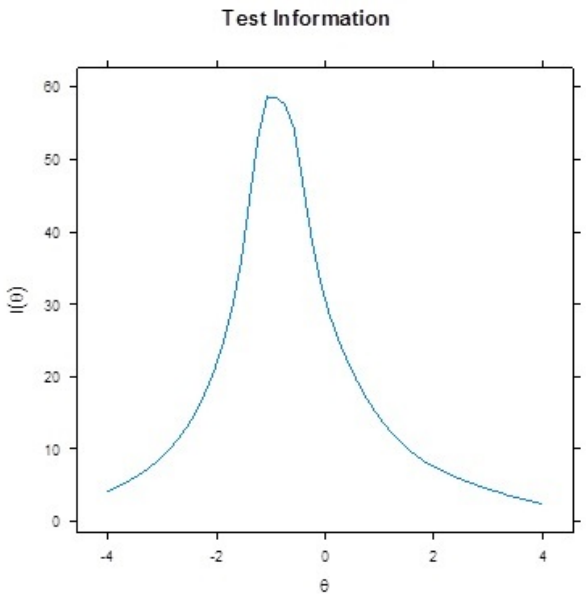

Figura 5.5: Curva de Informação do Teste.

5.3 Comparação entre as notas corrigidas pelo método Convencional e pelo Modelo de Resposta Gradual

A correção convencional padronizada das provas do vestibular da UnB elaboradas pelo CEBRASPE, como já citado anteriormente, é feita através da soma das pontuações adquiridas pelos alunos ao responderem as questões, subtraindo a média e dividindo pelo desvio padrão de todas as notas. É importante lembrar que o tipo de banco de dados abordado aqui se baseia na pontuação dos alunos adquirida através das respostas às questões do tipo A, a qual tem as seguintes características: se o aluno acertar a questão, ele ganha 1 ponto; se o aluno não responder a questão,

ele não recebe pontuação nenhuma; e por fim, se ele errar a questão, ele recebe a pontuação -1.

A diferença entre o cálculo das proficiências dos indivíduos pelo modelo da TRI e o cálculo da nota final pelo método convencional está no fato de que neste último a pontuação adquirida pelas respostas das questões é sempre a mesma, independentemente de se o respondente acertou uma questão fácil ou difícil, ou também, se errou uma questão fácil ou difícil. Isso ocorre pelo fato de que, apesar de essas questões terem graus de dificuldades diferentes, isso não é levado em consideração na nota final. Isto é, se um indivíduo acertar uma questão muito difícil, provando que ele tem um alto conhecimento sobre o traço latente medido, ele ganhará 1 ponto. Porém se outro indivíduo acertar uma questão muito mais fácil sobre esse mesmo traço latente, ele também receberá a mesma pontuação, 1. Já o modelo da teoria de resposta ao item funciona de forma distinta, pois há diferença na pontuação final do individuo caso ele acerte uma questão fácil ou uma questão difícil, ganhando menos pontos ou mais pontos, respectivamente.

Em específico, no modelo de resposta gradual, pode-se ter, ainda, uma diferença de pontuação da questão entre as três categorias de respostas, por exemplo: Em determinada questão, dependendo dos parâmetros de dificuldade de suas respectivas categorias, o indivíduo pode ganhar uma alta pontuação em acertar a questão e, caso ele erre ou não responda, pode não haver uma diferença tão grande entre a pontuação obtida ao se marcar uma dessas últimas categorias. Ou, também, pode-se ter uma questão em que se o indivíduo acertar ou não responder a questão, ele obtenha uma alta pontuação, e caso ele erre, obtenha uma pontuação bastante baixa e distante do valor das outras categorias. Assim, a distância entre as pontuações adquiridas a partir da escolha da categoria de cada questão vai variar, ou seja, não será padrão, pois no modelo de resposta gradual não é feita a soma das pontuações, mas sim a estimação das probabilidades baseado no modelo já mostrado no capítulo 2.

A partir do banco de dados simulado, foram feitas as correções convencional, que se dá através da soma das pontuações adiquiridas pelos alunos, e a convencional padronizada, que a partir das notas convencionais, subtraiu-se a média e dividiu-se pelo desvio padrão, dos 1000 indivíduos para as 50 questões. A análise descritiva das notas convencionais está apresentada na Tabela 5.5. Observa-se que, baseado

no número de questões, poderíamos esperar que tivéssemos um mínimo de -50, em que poderia haver algum aluno que iria errar todas as questões, e um máximo de 50, havendo algum aluno que acertasse todas as questões. Entretanto, podemos observar que obtivemos um mínimo de -25,00. Podemos atribuir este fato à consciência do aluno sobre a "vantagem" de deixar a questão em branco, fazendo com que, por exemplo, um indivíduo que não tenha estudado nada sobre o assunto, não chute todas as questões, mas sim, somente algumas, podendo ter errado todas, ou praticamente todas que respondeu, já que não tem domínio do assunto abordado.

Tabela 5.5: *Análise Descritiva das Notas Convencionais e das Notas pelo MRG (Dados Simulados).*

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Notas Convencionais	-25,0000	0,0000	5,0000	9,5880	16,0000	50,0000
Notas Convencionais Padronizadas	-2,8070	-0,7781	-0,3723	0,0000	0,5204	3,2800
Notas Modelo de Resposta Gradual	-1,8510	-0,7604	-0,3868	-0,0271	0,2995	3,6520

Na Figura 5.6 estão apresentadas as notas convencionais através de um histograma. Percebe-se uma alta frequência das notas em torno de zero e o restante da maioria dos valores das notas distribuídas acima de zero.

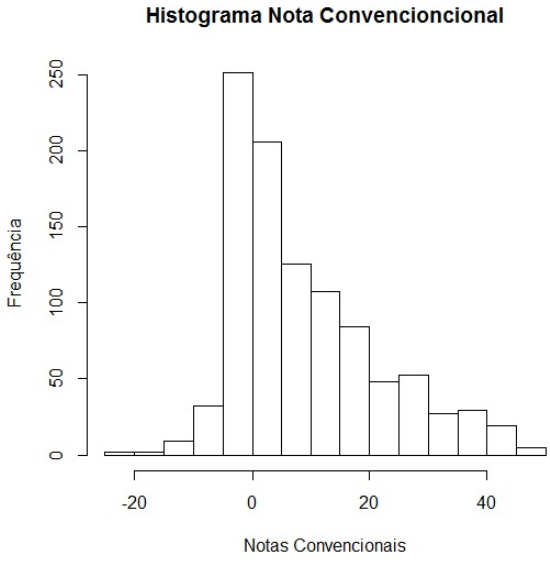

Figura 5.6: Histograma das Notas Convencionais.

Na Tabela 5.5, podemos observar também a análise descritiva das notas dos indivíduos corrigidas através do Modelo de Resposta Gradual. Observou-se um mínimo de -1,85 e um máximo de 3,65. Conforme apresentado no capítulo 2, no modelo de resposta gradual não se considera somente se o indivíduo respondeu a questão corretamente ou não, mas também qual foi a alternativa escolhida por ele, utilizando assim mais intensamente a informação contida no teste.

Dessa forma, não temos um parâmetro de mínimo e máximo das notas obtidas, pois cada categoria de cada item pode ter uma pontuação diferente, baseado na dificuldade de cada categoria do item e de sua respectiva discriminação.

Analisando ainda a Tabela 5.5, podemos observar que todas as medidas de posição das notas convencionais padronizadas estão com valores próximos dos valores obtidos através do MRG. Isso indica que os dois métodos geraram notas dos indivíduos similares. Na Figura 5.7, estão apresentadas graficamente as distribuições das notas convencionais padronizadas e das notas obtidas pelo MRG. Podemos observar que a maior concentração de notas ficou em torno de -1 até 0, com maior frequência de notas corrigidas pelo MRG, e o restante das notas se distribuem acima de 0.

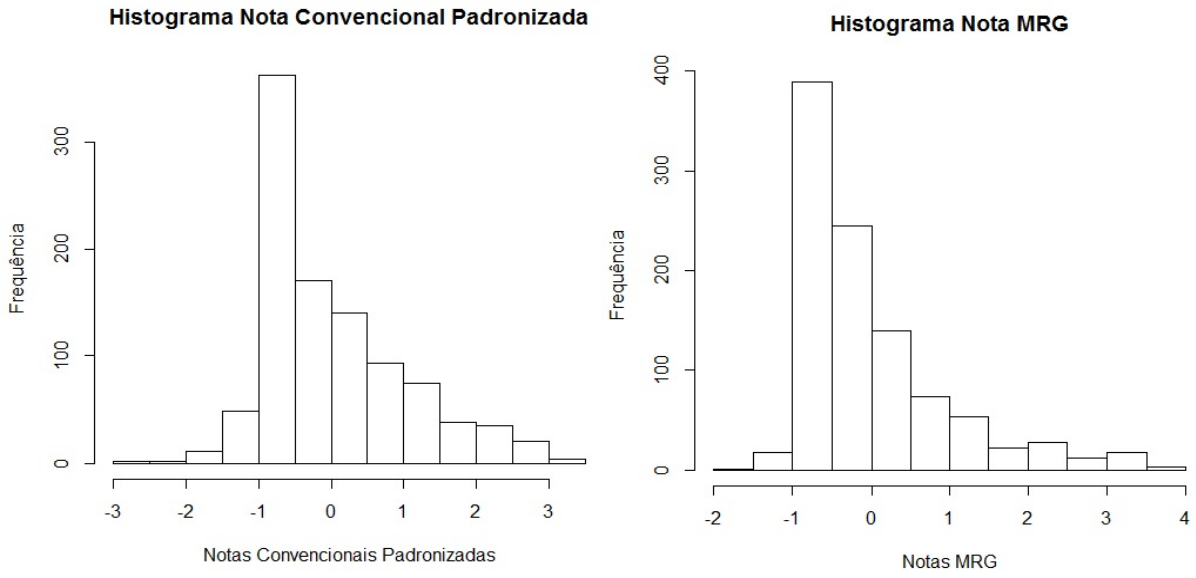

Figura 5.7: Histograma das Notas Convencionais Padronizadas e das Notas pelo MRG.

Para uma análise mais embasada, elaboramos um gráfico de correlação cruzando os valores das duas notas, em que as notas convencionais padronizadas estão apresentadas no eixo X e as notas pelo MRG no eixo Y, para analisarmos se há uma tendência linear entre as notas, indicando uma alta correlação entre seus respectivos valores.

Analisando esse gráfico (Figura 5.8), podemos perceber indícios de uma tendência linear comprovada pelo alto valor da correlação entre as notas, a qual resultou em 0,97562, indicando ainda que são diretamente proporcionais (Correlação Positiva). Sendo assim, há uma proximidade entre as notas. Entretanto, podemos observar que, para valores maiores (calda superior) ou menores das notas (calda inferior), há uma diferença entre os indivíduos, fazendo que com os dados percam um pouco a linearidade.

Essas diferenças na calda superior, por exemplo, indicam, talvez, que os indivíduos acertaram questões mais difíceis, não fazendo diferença no método convencional, porém, o modelo de resposta gradual levou em conta a dificuldade desses itens, fazendo com que eles passem a ter uma nota maior, o que não ocorre no método convencional. Já na calda superior, os indivíduos passam a ter também uma nota um pouco maior, pelo fato de que, talvez, eles possam não ter respondido questões fáceis/difíceis, por exemplo, ganhando uma pontuação mínima no modelo de resposta gradual, ao não responderem a questão, aumentando sua nota, ao invés de simplesmente não ganharem nada, como no modelo convencional.

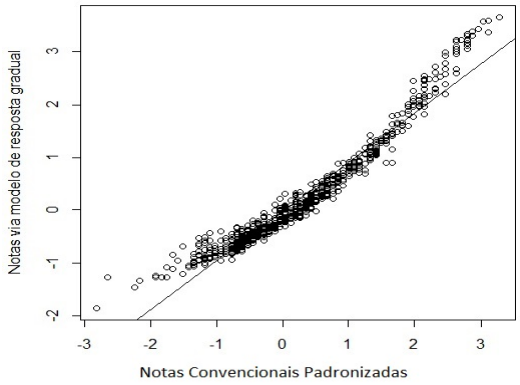

Figura 5.8: Diagrama de Dispersão entre as Notas Convencionais Padronizadas e as Notas pelo MRG.

Avaliando alguns casos em que a diferença de notas entre os dois métodos foi significativa (foi definida uma diferença de 0,20 ou mais como significativa, apenas a título de comparação descritiva), observamos que para o indivíduo 16, por exemplo, houve uma mudança de nota de 0,8012 para 0,5720, obtida pelo método convencional e pelo MRG, respectivamente. Analisando as respostas deste individuo (Tabela 5.6) para cada questão do teste e associando aos valores dos parâmetros de dificuldade de cada item, observamos, que de fato, em geral, o aluno só acertou questões consideradas fáceis, e mesmo assim, errou algumas, e não respondeu questões consideradas mais difíceis, fazendo sentido sua nota diminuir, já que o conhecimento dele não seria tão grande sobre o assunto abordado.

Por outro lado, analisando o indivíduo 13, por exemplo, observou-se que este aumentou de nota de 2,8741 para 3,2512, obtida pelo método convencional e pelo MRG, respectivamente. Através da análise das respostas deste individuo e dos valores dos parâmetros dos itens, podemos perceber indícios de que também a nota do MRG poderia ser considerada mais "justa", pois observa-se que o indivíduo deixou somente duas questões em branco, sendo estas questões consideradas as mais difíceis, e respondeu todas as outras questões, acertando a maioria delas, principalmente as com dificuldades relevantes. Dessa forma, podemos concluir que o aluno é avaliado de forma mais criteriosa e "justa" através da correção pelo MRG. Esses valores estão demonstrados na Tabela 5.6.

Tabela 5.6: *Respostas dos Indivíduos 16 e 13. (Dados Simulados).*

Questão	a_i	b_{i1}	b_{i2}	Respostas Indivíduo 16	Respostas Indivíduo 13
1	1,615	-1,479	-0,207	1	1
2	1,620	-1,555	-0,333	1	1
3	5,145	-1,143	-0,731	1	1
4	0,907	-1,858	0,501	1	1
5	3,861	-1,146	-0,558	1	1
6	2,481	-1,226	-0,344	1	1
7	4,581	-1,101	-0,570	1	1
8	4,388	-1,169	-0,617	1	1
9	2,474	-1,282	-0,361	1	1
10	3,083	-1,276	-0,521	1	1
11	1,330	-1,609	0,177	-1	1
12	1,898	-1,472	-0,178	1	1
13	1,057	-1,748	0,623	1	1
14	1,076	-1,825	0,471	1	-1
15	2,354	-1,317	-0,124	1	1
16	1,795	-1,409	0,104	1	1
17	2,573	-1,234	-0,117	1	1
18	1,793	-1,447	0,117	1	1
19	2,988	-1,269	-0,116	1	1
20	1,041	-1,812	0,840	-1	1
21	2,777	-1,384	-0,093	1	1
22	1,910	-1,512	0,251	1	1
23	2,249	-1,437	0,148	1	1
24	1,604	-1,608	0,514	1	1
25	1,361	-1,654	0,798	1	1
26	0,779	-2,294	1,693	0	1
27	2,033	-1,565	0,460	1	1
28	1,770	-1,607	0,711	-1	1
29	1,404	-1,826	0,981	-1	1
30	1,960	-1,692	0,717	-1	1
31	0,684	-2,678	2,359	0	1
32	1,110	-2,132	1,538	0	1
33	1,726	-1,985	1,023	0	1
34	1,208	-2,115	1,627	0	1
35	0,489	-3,411	3,836	0	1
36	1,175	-2,471	1,794	0	1
37	1,143	-2,467	1,918	0	1
38	1,134	-2,683	2,037	0	1
39	0,143	-12,215	14,579	0	1
40	0,921	-3,115	2,608	0	1
41	0,981	-3,198	2,535	0	1
42	0,663	-4,020	4,115	0	1
43	0,795	-3,510	3,549	0	-1
44	0,637	-4,586	4,509	0	1
45	0,360	-6,749	7,854	0	1
46	0,897	-3,886	3,518	0	1
47	0,778	-4,338	4,166	0	1
48	0,207	-14,141	15,081	0	1
49	0,024	-104,808	129,833	0	0
50	-0,054	44,041	-57,349	0	0

Com o intuito de confirmar a relação entre as notas convencionais padronizadas e as notas obtidas pelo modelo de reposta gradual, desconsiderando qualquer suposição sobre esta relação, realizamos um ajuste dos dados utilizando uma técnica estatística não paramétrica, chamada Regressão de Nadaraya-Watson. Essa teoria será abordada, de forma breve, a seguir.

5.3.1 Regressão de Nadaraya-Watson

De forma geral, os principais objetivos da modelagem estatística Regressão é descrever o relacionamento entre uma ou mais variáveis explicativas, independentes, e uma determinada variável de interesse, dependente. Para isso, diferentes tipos de regressões paramétricas e não paramétricas foram desenvolvidas. A principal diferença entre essas análises é que a regressão paramétrica assume que a relação funcional entre a variável resposta e as preditoras é conhecida, ou seja, se é estabelecida uma suposição a priori de como é essa relação. Já na regressão não paramétrica, sua característica principal é a ausência (completa ou quase completa) de conhecimento a priori a respeito da forma funcional que está sendo estimada.

Dentre os métodos não paramétricos de estimação de curvas de regressão, podemos citar a Regressão de Nadaraya-Watson, a qual é comumente utilizada para estimar uma função de regressão quando não se faz nenhuma suposição a priori sobre a mesma. Quando se deseja ajustar uma função que represente a relação entre Y, variável dependente, e X, uma variável independente, de forma que

$$Yi = m(x_i) + \varepsilon_i, i = 1, 2, ...n \tag{5.1}$$

deve-se estimar $m(x_i)$, desconhecida, sem restringi-la a modelos estabelecidos a priori, como ocorre no caso paramétrico.

Nadaraya (1964) e Watson (1964) desenvolveram uma forma de estimar esta função, dando origem ao estimador de Nadaraya-Watson, dado por:

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_x(\frac{x - x_i}{h_n})y_i}{\sum_{i=1}^{n} K_x(\frac{x - x_i}{h_n})} \tag{5.2}$$

em que h_n é o parâmetro de suavização, que é uma espécie de controlador entre o vício e a variância da estimativa, e K_x é uma função densidade de probabilidade, consistindo o método em estabelecer uma média localmente ponderada. Como detalhar essa teoria não é o foco principal deste trabalho, para um estudo mais detalhado sobre esta técnica, consultar Silverman(1996), Cacoullos (1996), Simonoff(1996), Hastie e Tibshirani(1990), Cortes(2010), Silva(2010), Simonassi e Júnior (2005), Hunter(2001), Cai(2011).

A título de exemplificação, a Figura 5.9 apresenta um exemplo da comparação de uma curva m (curva representando a disposição dos pontos no gráfico), a qual pretende-se estimar, e a curva aproximada, utilizando o estimador de Nadaraya-Watson.

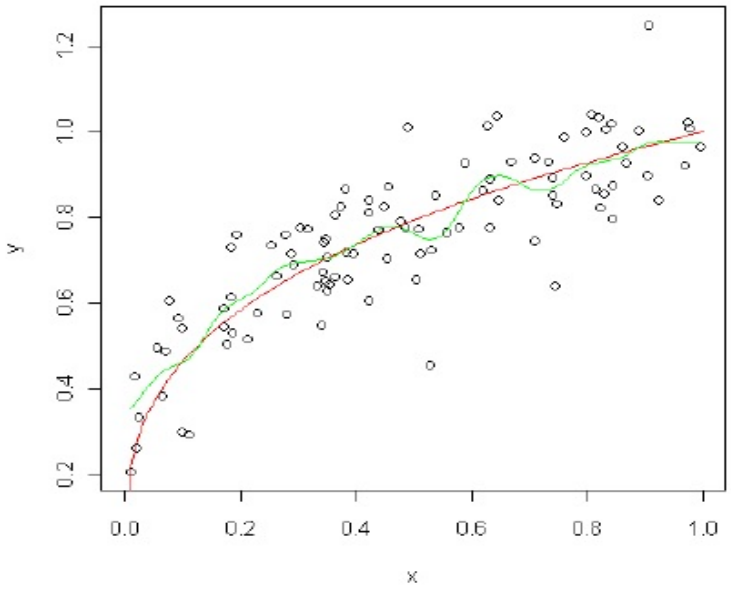

Figura 5.9: Exemplo da curva m (vermelho) e da curva estimada por Nadaraya-Watson. Fonte: Silva (2010).

Conforme apresentado na Figura 5.8, observamos indícios de que as notas obtidas pelo modelo de resposta gradual e pelo método convencional tenham uma relação de linearidade, diretamente proporcional, mostrando assim uma proximidade entre as notas dos indivíduos ao compararmos os dois métodos.

Analisando agora a ordem de classificação destes indivíduos, ajustamos o modelo de Nadaraya-Watson entre o rank obtido pelo método convencional e o rank obtido pelo MRG. Na Figura 5.10, em que o eixo horizontal corresponde ao rank Con-

vencional e o eixo vertical corresponde ao rank do MRG, está apresentado o resultado do ajuste da curva através da Regressão de Nadaraya-Watson, e podemos observar que há também uma tendência linear entre os ranks, com um alto grau de conformidade entre os indivíduos que tiraram notas maiores, obtendo uma boa classificação no rank (canto inferior esquerdo do gráfico), e um grau de conformidade menor entre os indivíduos que obtiveram notas menores, ocupando as últimas colocações (canto superior direito do gráfico).

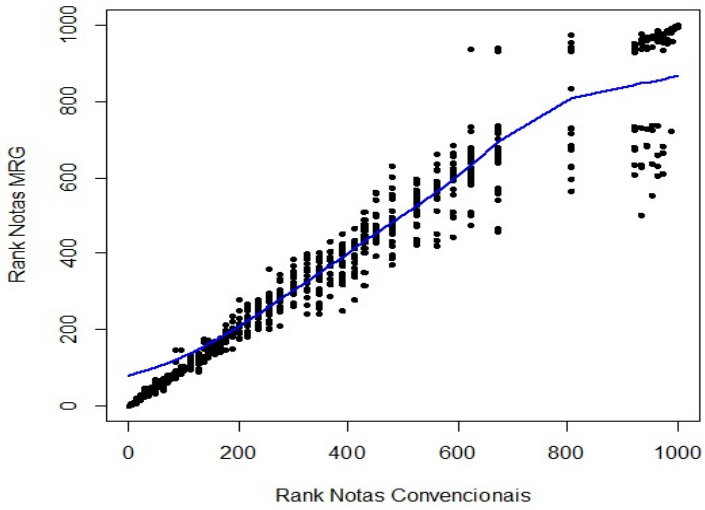

Figura 5.10: Curva estimada pelo método de Nadaraya-Watson entre os Ranks obtidos pelo método MRG e pelo método Convencional (Dados Simulados).

É possível perceber de forma mais clara essa diferença entre os ranks dos indivíduos com maior ou menor classificação na Figura 5.11, em que no eixo horizontal está plotado o rank convencional e no eixo vertical está plotado o módulo da diferença entre os ranks convencional e MRG, também obtido pelo ajuste de Nadaraya-Watson. Podemos observar que, de fato, a diferença é bem menor entre os indivíduos que ocupam as primeiras colocações, ao compararmos os ranks obtidos pelos indivíduos com menor classificação.

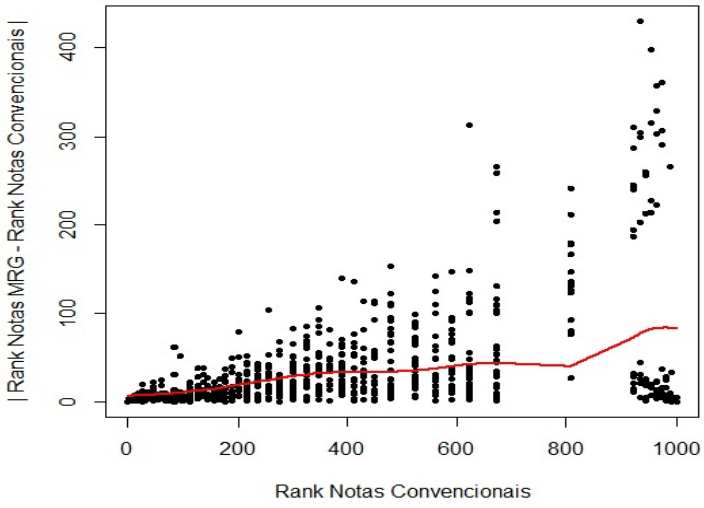

Figura 5.11: Curva estimada pelo método de Nadaraya-Watson entre o módulo da diferença entre os Ranks obtidos pelo método MRG e pelo método Convencional e Rank Convencional (Dados Simulados).

Dessa forma, podemos concluir que a correção das questões de uma prova pelo modelo de resposta gradual tem a mesma eficiência em estimar as notas dos indivíduos que o método convencionalmente utilizado, gerando notas aproximadas para os indivíduos em colocações mais altas, sendo estes os que serão relevantes em uma avaliação, pois o número de vagas é limitado. Entretanto, apesar de haver indícios de calcular notas semelhantes para os indivíduos, o modelo de resposta gradual da TRI tem algumas vantagens sob o método convencional, tais como: possível comparação entre populações diferentes, submetidos a provas diferentes com algumas questões em comum, avaliação mais "justa" do traço latente medido, levando em consideração a dificuldade e discriminação do item, entre outras vantagens.

5.4 Outros Contextos

Com o intuito de compararmos não somente as notas propriamente ditas dos indivíduos, mas também a diferença entre as classificações dos candidatos obtidas pelos dois métodos de correção, o convencional e o MRG, foram construídos gráficos em que foram plotados os níveis de discordância entre ranks obtidos pelos dois métodos

versus o número de candidatos selecionados, para cada combinação de casos variando o número de itens (50,100, 150 e 200 itens) e indivíduos (1000, 5000, 10000 e 20000).

O nível de discordância foi calculado para cada número de vagas. Na abscissa (eixo horizontal), está o número de vagas. Dentro deste número de vagas, o nível de discordância será a proporção de selecionados pelo modelo de resposta gradual que não foram selecionados pelo método convencional. Por exemplo, para o caso de 20000 candidatos, o valor da ordenada (eixo vertical) para o valor 1000 na abscissa corresponde à proporção de discordância entre os 1000 primeiros candidatos pelas duas notas, isto é, para um nível de discordância de 0.04, significa que 4% dos 1000 primeiros colocados pela nota convencional padronizada não estão entre os 1000 primeiros colocados pela nota obtida via MRG. Isto é recalculado para cada número de vagas, de 2 até 20000. Claramente, podemos constatar que para 20000 vagas, o nível de discordância tem que ser zero.

Com o objetivo de retratar essa relação dos níveis de discordância entre os ranks e o número de candidatos, para cada combinação de variação de indivíduos e itens, foram geradas 100 amostras diferentes, gerando assim bancos de dados de respostas e ranks das notas dos indivíduos obtidos pelos dois métodos diferentes. Após isso, foi estimada uma curva média das amostras dos dados simulados para cada caso de número de indivíduos e número de itens, representando assim a relação média entre o nível de discordância entre os ranks em função do número de candidatos. Além disso, foi estimada também essa curva média através da regressão de Nadaraya-Watson, confirmando assim, o comportamento da curva apresentada.

Para o primeiro caso, em que corresponde ao banco de dados gerado com 1000 indivíduos e 50 itens (Figura 5.12), podemos perceber através da curva média (linha mais clara) que, em média, o grau de discordância é abaixo de 5% em torno 150 candidatos, aproximadamente. Isso indica que se, por exemplo, a prova do vestibular da UnB fosse composta somente por esse tipo de questão considerada (questões do tipo A) e os candidatos tivessem concorrendo a 150 vagas, aproximadamente somente 7 indivíduos, dos 150 primeiros colocados, que estariam dentro do número de vagas quando avaliados pelo método convencional, não estariam dentro do número de vagas quando avaliados pelo método MRG.

Podemos observar ainda, que esse baixo grau de discrepância se matém até aproximadamente 600 vagas. A partir de 600 vagas até em torno de 800 vagas, podemos verificar um alto grau de discrepância para os indivíduos que ficaram nesta colocação do rank, em decorrência de um alto grau de empate nas notas mais baixas. Para os indivíduos que ficaram nas últimas classificações, como esperado, obtemos um grau de discordância tendendo a zero.

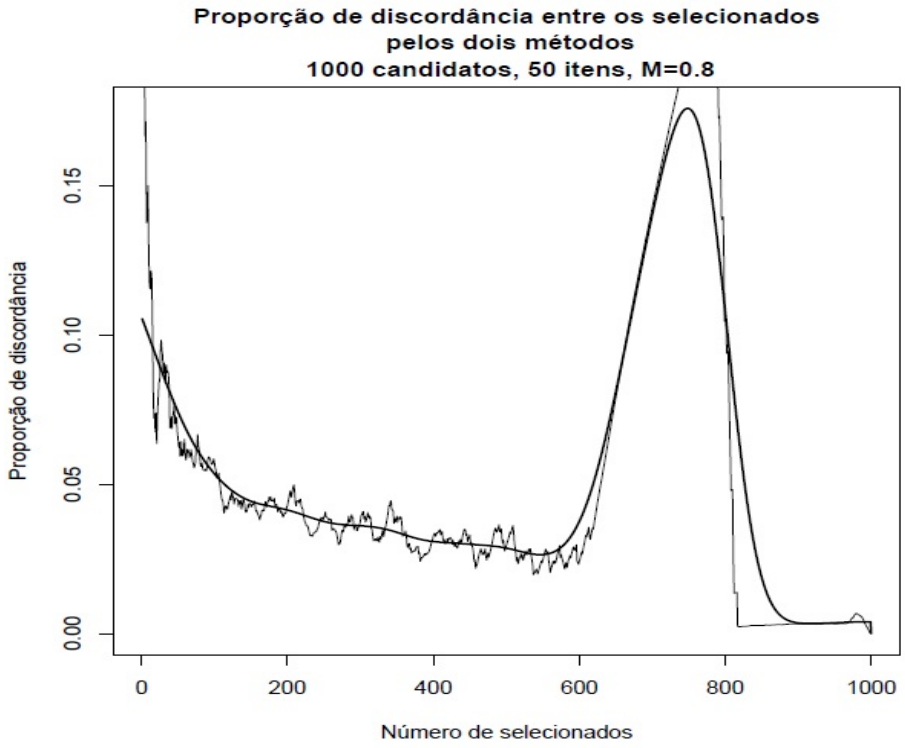

Figura 5.12: Gráfico de discordância entre o método convencional e MRG. (1.000 indivíduos e 50 itens - Dados Simulados)

A partir dos gráficos que representam os próximos três casos (Figuras 5.13, 5.14 e 5.15), os quais foram gerados também com 1000 indivíduos, alterando somente a quantidade de itens de 50 para 100, 150 e 200 itens, podemos observar que o grau de discrepância resulta em menos de 5% para uma quantidade menor de vagas, ao se comparar com o gráfico anteriormente analisado.

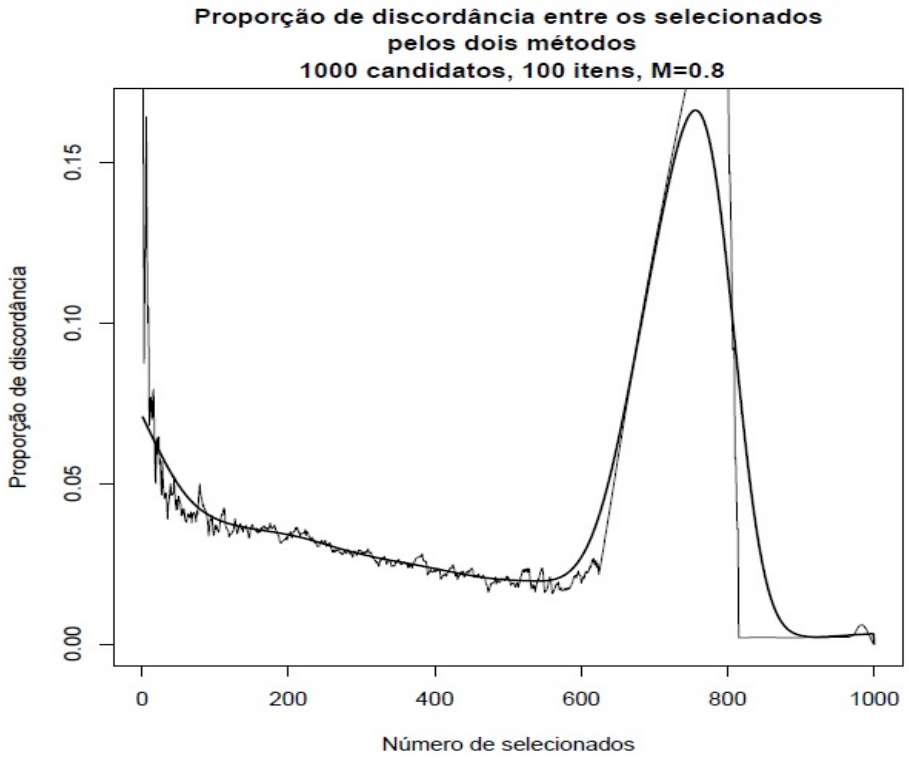

Figura 5.13: Gráfico de discordância entre o método convencional e MRG. (1.000 indivíduos e 100 itens - Dados Simulados)

Por exemplo, para o gráfico que corresponde a 100 itens respondidos (Figura 5.13), em aproximadamente 100 vagas, o grau de discrepância já resulta abaixo de 5%, e para os gráficos seguintes (Figuras 5.14 e 5.15), com 150 e 200 itens, esse grau de discordância é ainda menor para a mesma quantidade de vagas. Em relação às curvas médias estimadas através da regressão de Nadaraya-Watson (linha mais escura), podemos perceber que há uma concordância entre as curvas, indicando uma boa estimação do comportamento dos dados de forma suavizada.

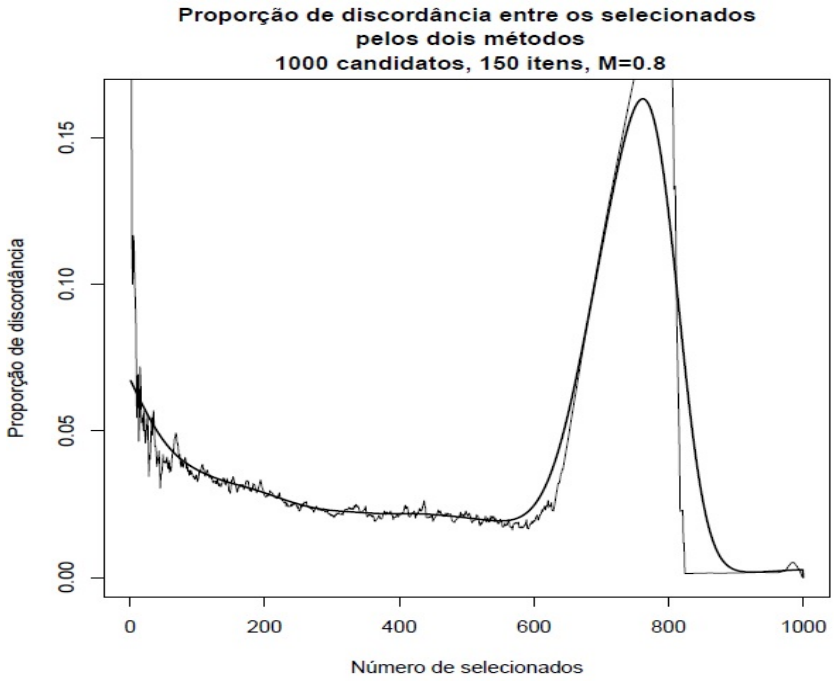

Figura 5.14: Gráfico de discordância entre o método convencional e MRG. (1.000 indivíduos e 150 itens - Dados Simulados)

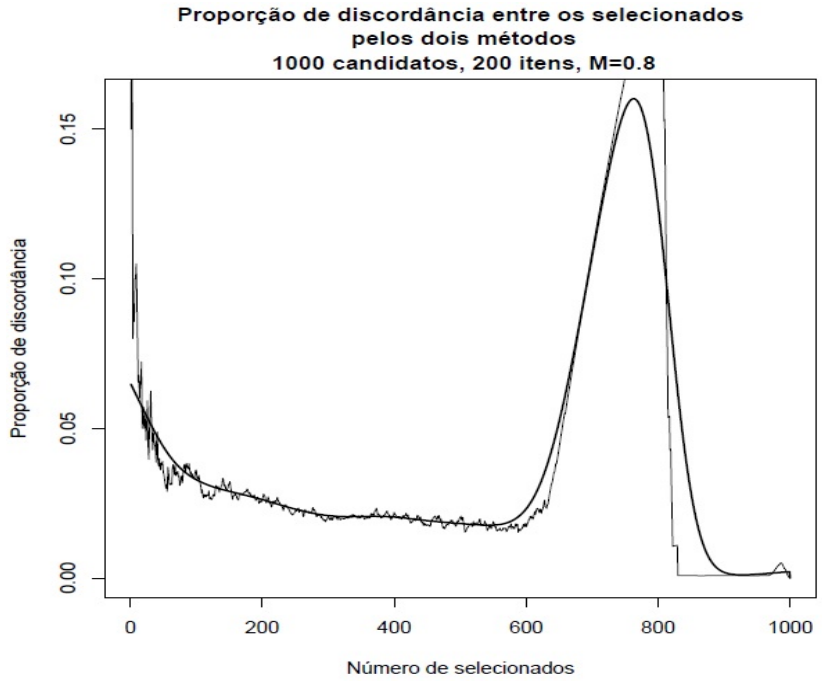

Figura 5.15: Gráfico de discordância entre o método convencional e MRG. (1.000 indivíduos e 200 itens - Dados Simulados)

Analisando agora o caso gerado com 5000 indivíduos e 100 itens (Figura 5.16), podemos perceber que no ponto que representa 1000 indivíduos aproximadamente, o grau de discordância é em torno de 4%, indicando que, por exemplo, para um número limite de vagas para 1000 candidatos, 40 indivíduos que seriam classificados pelo método de correção convencional, não seriam classificados pelo método de correção do modelo de resposta gradual. Podemos observar, também, o mesmo comportamento da curva dos casos discutidos anteriormente para indivíduos que tiveram notas baixas, classificados no ponto em torno de 3000 a 4000 na abscissa. E também, da mesma forma, podemos perceber que para os últimos candidatos o grau de concordância tende a zero. A curva suavizada pela regressão de Nadaraya-Watson resultou em uma alta concordância com a curva média estimada.

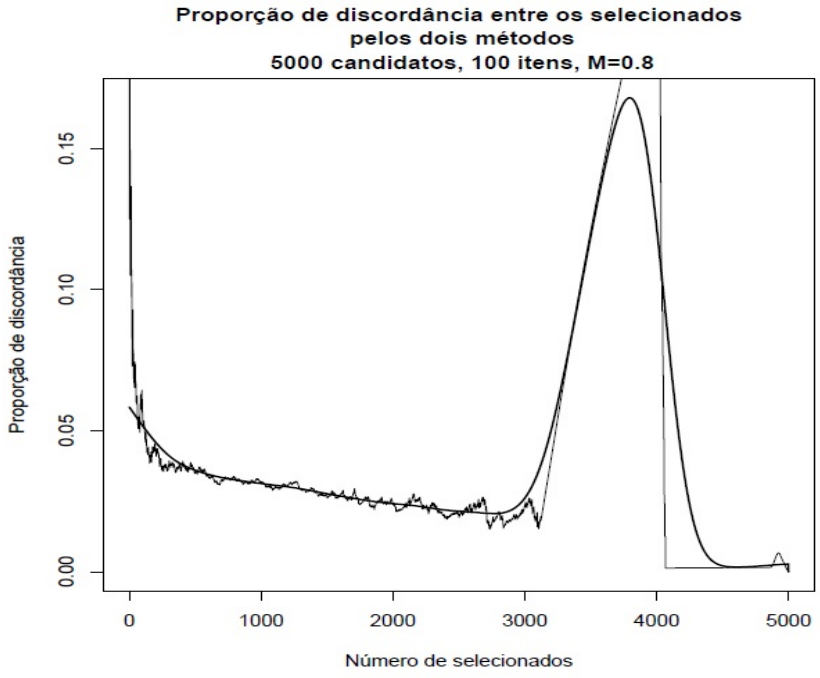

Figura 5.16: Gráfico de discordância entre o método convencional e MRG. (5.000 indivíduos e 100 itens - Dados Simulados)

Para os casos com 5000 indivíduos e 150 e 200 itens (Figuras 5.17 e 5.18), podemos perceber também um comportamento similar, em que o número de discrepância entre os ranks obtidos pelos dois métodos é considerado baixo a partir de 1000 candidatos até em torno de 3000 candidatos.

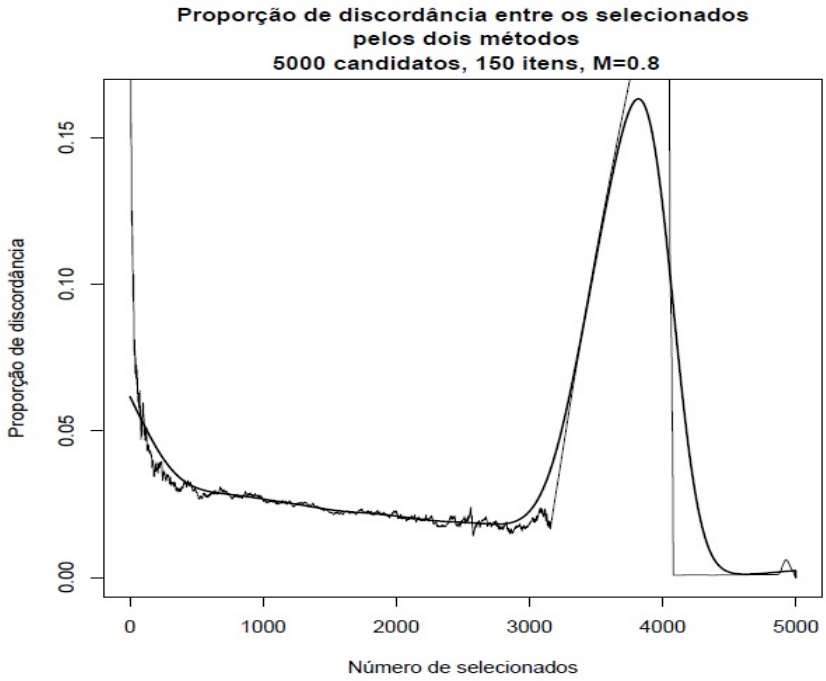

Figura 5.17: Gráfico de discordância entre o método convencional e MRG. (5.000 indivíduos e 150 itens - Dados Simulados)

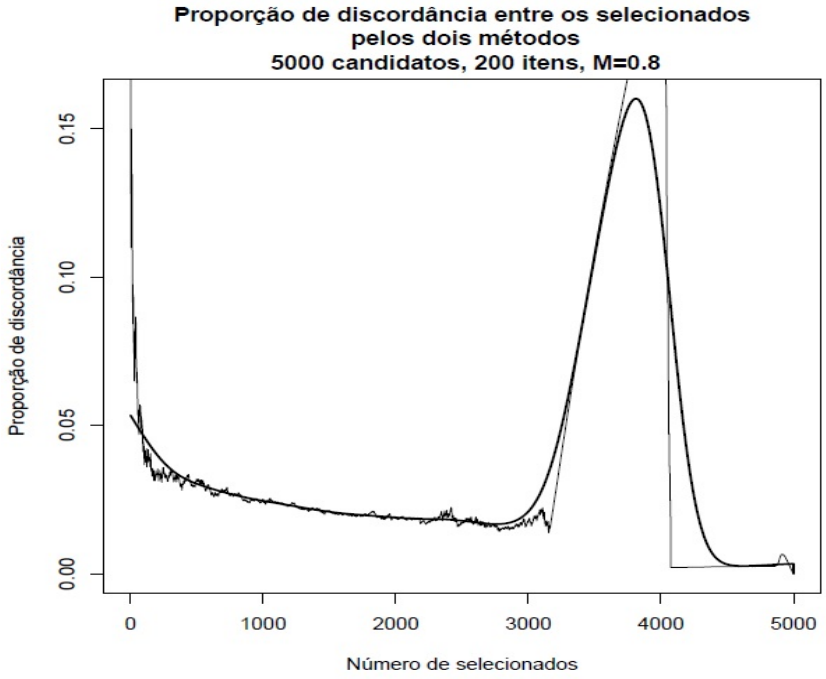

Figura 5.18: Gráfico de discordância entre o método convencional e MRG. (5.000 indivíduos e 200 itens - Dados Simulados)

Já para o gráfico em são considerados 5000 indivíduos e 50 itens (Figura 5.19), podemos perceber que o grau de discordância oscila muito quando o número de selecionados está entre 1000 e 3000 candidatos. Podemos atribuir esse comportamento ao fato de serem poucos itens para avaliar muitos indivíduos.

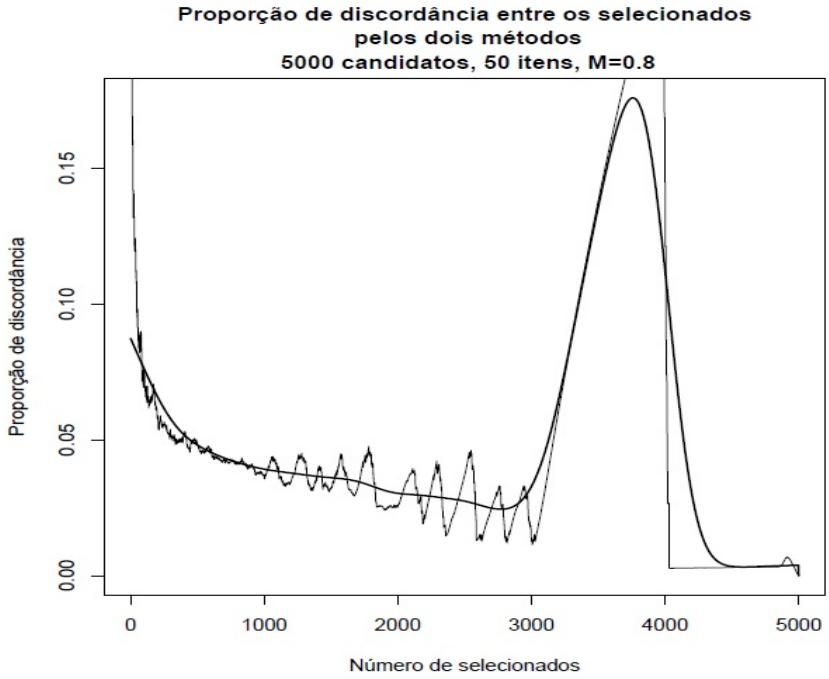

Figura 5.19: Gráfico de discordância entre o método convencional e MRG. (5.000 indivíduos e 50 itens - Dados Simulados)

Analisando agora os casos com 10.000 indivíduos, percebemos que o grau de discrepância entre as classificações oriundas dos dois métodos de correção, no gráfico gerado com 150 questões (Figura 5.20), é alto para poucas vagas, proporcionalmente, e começa a se estabilizar em torno de 1000 a 2000 mil vagas. Esse comportamento se mantém até aproximadamente 6000 mil vagas e em seguida há o pico de alto grau de discordância, também já detectado nos outros gráficos, para os participantes que obtiveram notas mais baixas. Podemos perceber um comportamento similar para o gráfico gerado com 200 itens (Figura 5.21).

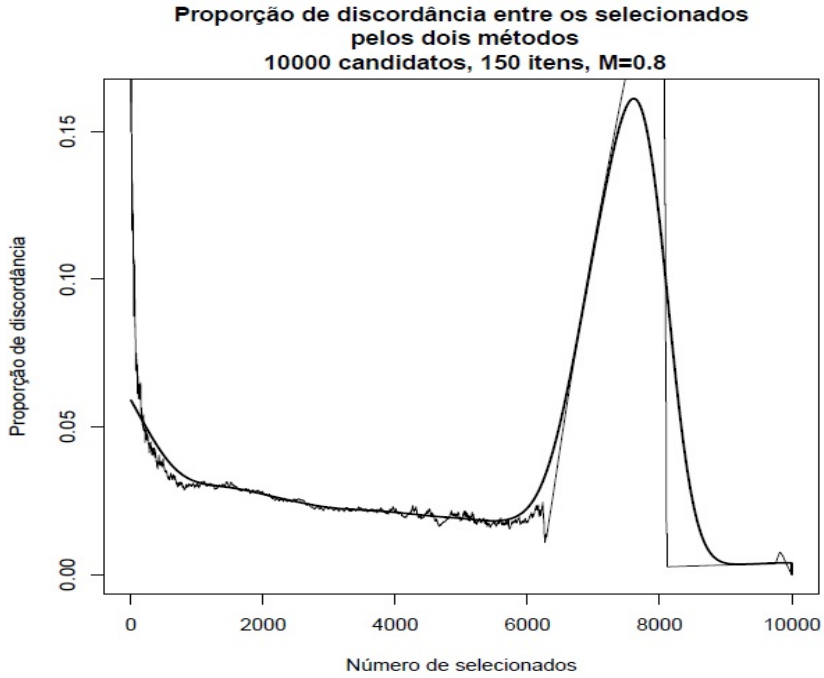

Figura 5.20: Gráfico de discordância entre o método convencional e MRG. (10.000 indivíduos e 150 itens - Dados Simulados)

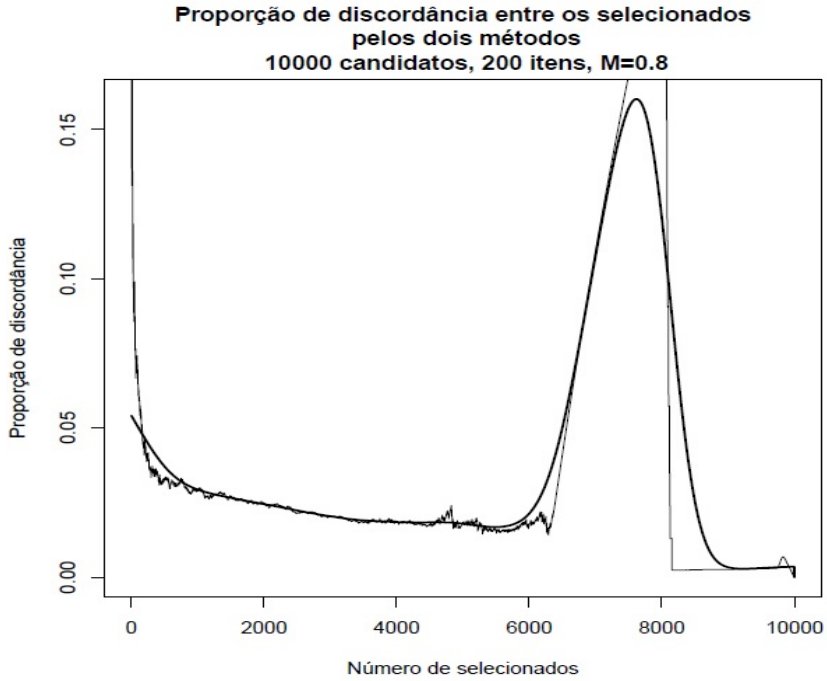

Figura 5.21: Gráfico de discordância entre o método convencional e MRG. (10.000 indivíduos e 200 itens - Dados Simulados)

Ainda sobre os casos com 10.000 indivíduos, no gráfico correspondente à análise de 50 itens (Figura 5.22), observamos um mau comportamento da curva a partir de 2000 candidatos, até aproximadamente 6000. No gráfico correspondente a 100 itens (Figura 5.23) isso também ocorre, entretanto, somente a partir de aproximadamente 4000 candidatos, resultando em um intervalo de oscilação menor.

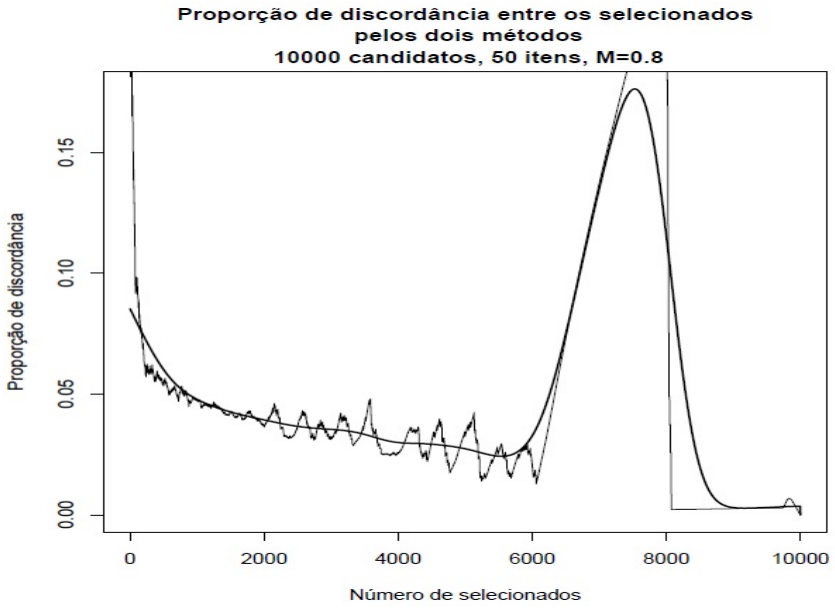

Figura 5.22: Gráfico de discordância entre o método convencional e MRG. (10.000 indivíduos e 50 itens - Dados Simulados)

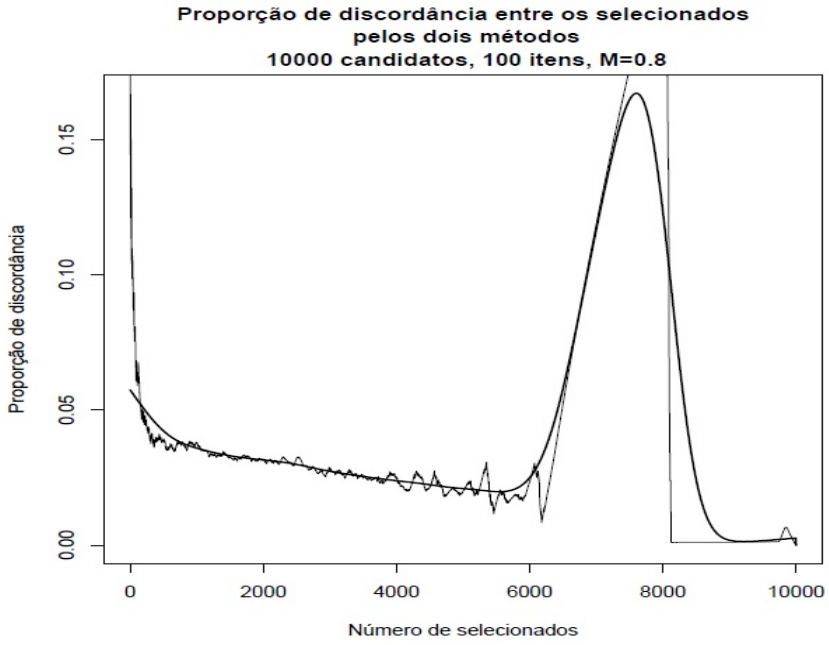

Figura 5.23: Gráfico de discordância entre o método convencional e MRG. (10.000 indivíduos e 100 itens - Dados Simulados)

Analisando agora os casos em que temos 20.000 indivíduos e as suas respectivas variações de número de itens (50, 100,150 e 200), temos na Figura 5.24 o gráfico em que foram considerados 200 itens. Podemos perceber um comportamento semelhante aos casos citados anteriormente, nas combinações proporcionais de crescimento entre o número de indivíduos e o número de itens considerados. Esse comportamento é caracterizado pela situação em que, quando há um número muito baixo de indivíduos selecionados, temos um alto grau de discrepância, mas à medida que o número de candidatos selecionados aumenta, esse grau de disparidade entre os ranks obtidos pelos dois métodos de correções tende a diminuir e se manter baixo até, aproximadamente, 13.000 candidatos, havendo um pico de alta discordância, se mantendo até próximo de 16.000 candidatos, e depois esse grau tendendo a zero.

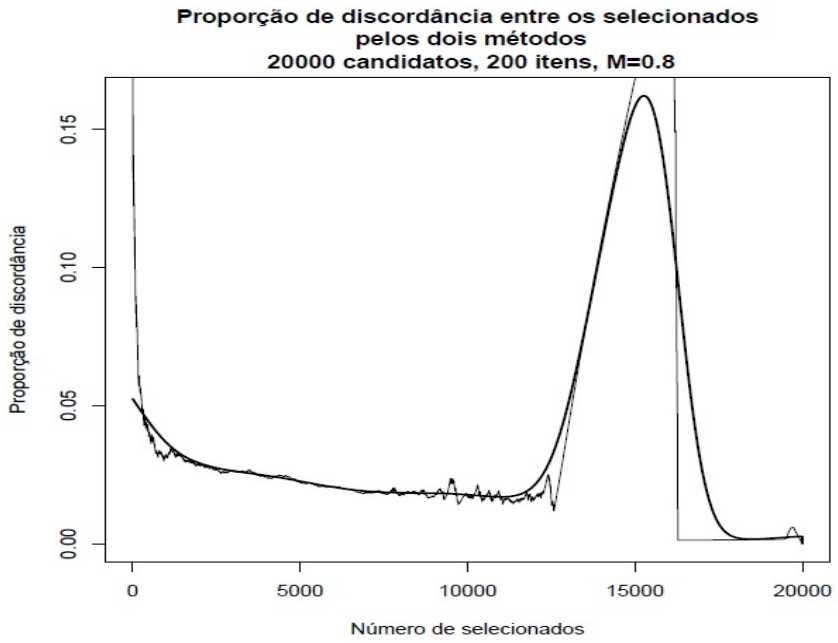

Figura 5.24: Gráfico de discordância entre o método convencional e MRG. (20.000 indivíduos e 200 itens - Dados Simulados)

Em específico, podemos perceber que num total de aproximadamente 1000 vagas, entre esses 20.000 participantes, o grau de discordância resultou em torno de 0,04, indicando que 4% dos 1000 primeiros indivíduos, ou seja, 40 candidatos, seriam selecionados pelo método convencional, mas não seriam selecionados pelo MRG. Dessa forma, os 960 primeiros indivíduos selecionados pelo método convencional seriam também selecionados pelo método MRG, e os outros 40 candidatos selecionados pelo MRG estariam além dos 1000 primeiros indivíduos selecionados pelo método convencional.

Consideramos essa situação a que mais se aproxima de uma situação real do vestibular da UnB, em que participam da seleção do vestibular em torno de 20.000 indivíduos que são submetidos a aproximadamente 200 questões do tipo A. Se, por acaso, as provas do vestibular fossem compostas somente por este tipo de questão, e se, por exemplo, fossem selecionados 4.000 candidatos, poderíamos concluir que somente, aproximadamente 3% dos candidatos, ou seja, 120 candidatos estariam classificados de forma diferente entre os métodos em questão. Podemos atribuir esse fato, em que estes 120 indivíduos seriam classificados somente pelo método MRG e não pelo método convencional, ao pressuposto de que estes tenham acertado questões

consideradas difíceis pelo método MRG, efeito esse que o método convencional não consegue capturar. Além disso, podemos perceber ainda uma boa estimação pela Regressão de Nadaraya-Watson. Analisando a Figura 5.25, em que são considerados 20.000 indivíduos e 150 itens, percebemos também um comportamento parecido com o gráfico analisado anteriormente.

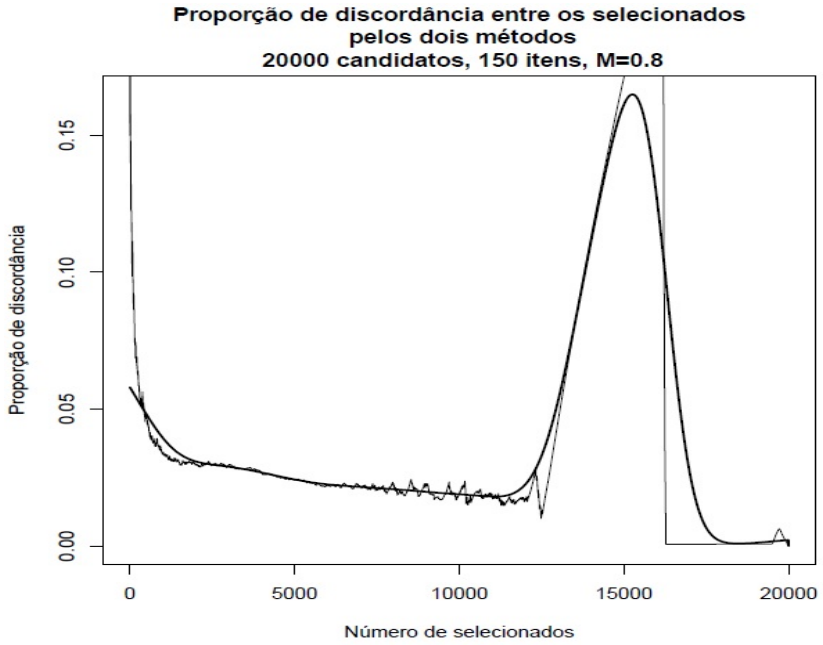

Figura 5.25: Gráfico de discordância entre o método convencional e MRG. (20.000 indivíduos e 150 itens - Dados Simulados)

Analisando agora as Figuras 5.26 e 5.27, em que são considerados 100 itens e 50 itens respectivamente, e com 20.000 indivíduos, observamos a oscilação já detectada anteriormente, quando o número de indivíduos avaliados é muito maior do que o número de itens, nas combinações de indivíduos e itens consideradas neste trabalho. Observamos esse comportamento não constante da curva começando a partir de aproximadamente 5000 indivíduos até em torno de 13.000, sendo que, no último gráfico, esse intervalo é um pouco maior, começando em aproximadamente 4000 candidatos.

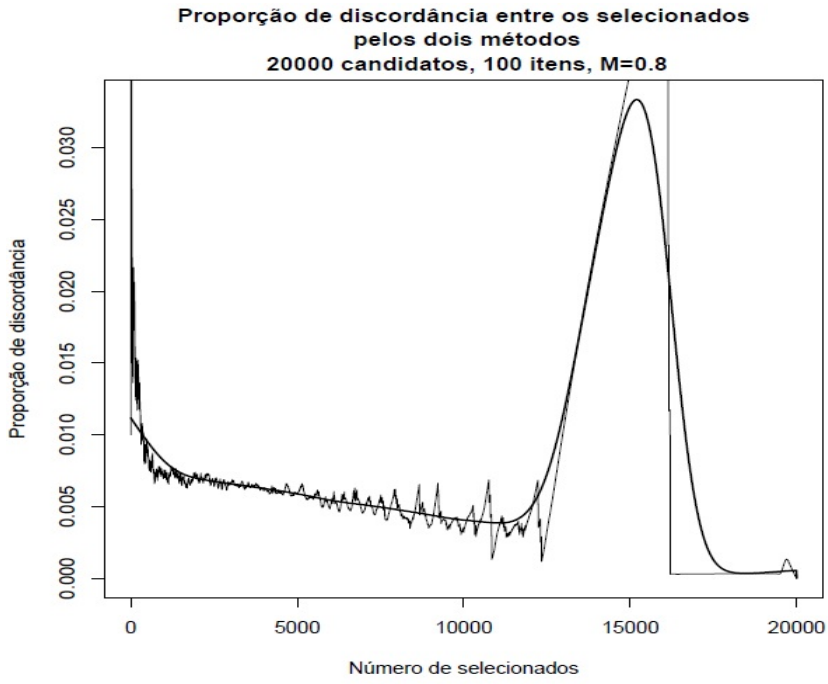

Figura 5.26: Gráfico de discordância entre o método convencional e MRG. (20.000 indivíduos e 100 itens - Dados Simulados)

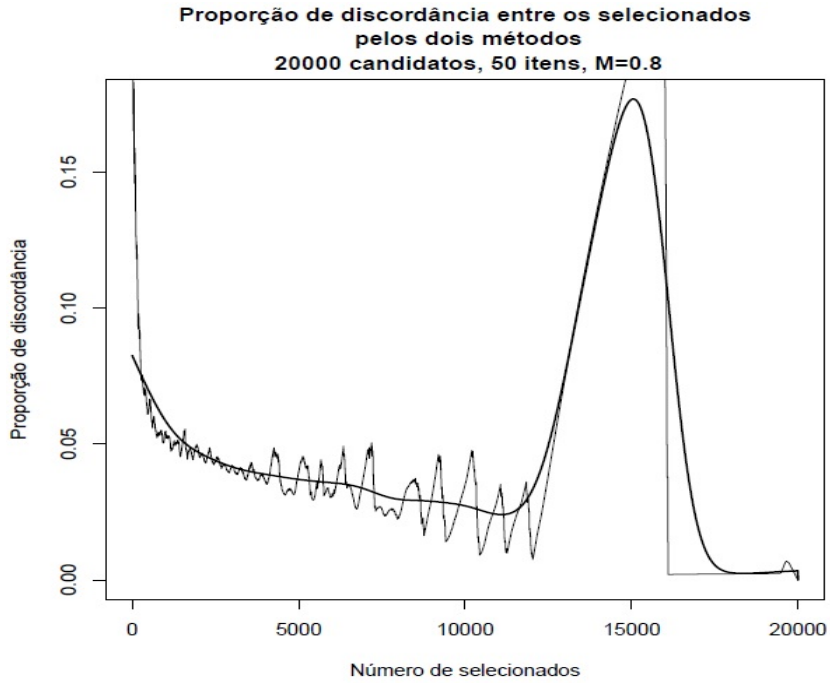

Figura 5.27: Gráfico de discordância entre o método convencional e MRG. (20.000 indivíduos e 50 itens - Dados Simulados)

Capítulo 6

Resultados- Dados Reais

Neste capítulo, serão apresentados os resultados obtidos a partir das análises do banco de dados fornecido pelo CESBRASPE, conforme descrito na seção 4.3. Para a manipulação da base de dados cedida pelo CEBRASPE, foi utilizado o Software SAS. Da mesma forma que nos dados simulados, descritos no capítulo anterior, o programa utilizado para gerar os resultados foi o Software R e as funções utilizadas para gerar o modelo de resposta gradual, calcular os escores dos indivíduos e estimar os parâmetros dos itens foram as funções *mirt*, *fscores* e *coef*, respectivamente.

6.1 Análise Descritiva

De acordo com o que foi apresentado na seção 4.3, foi disponibilizado pelo CEBRASPE um banco de dados contendo as respostas de 21.969 alunos às 300 questões aplicadas no vestibular da UnB de 2014/2. Pelos motivos já descritos na seção citada, nesta análise foram usadas somente as respostas de 7232 indivíduos de 237 questões. Essas questões referiam-se as seguintes áreas: Língua Portuguesa, História, Geografia, Artes Visuais, Música, Filosofia, Antropologia, Biologia, Física, Matemática e Química.

Após a manipulação da base de dados do vestibular da UnB para que a base ficasse em um formato propício para análise, foi avaliada a frequência de respostas por indivíduo de forma geral, entre as opções de categorias: Errar (-1), Não responder (0) e Acertar (1).

Nas Tabelas 6.1 a 6.4, podemos observar diferentes perfis de alunos, agora baseados na realidade de fato, ao contrário da situação descrita no capítulo anterior, o qual mostra as análises dos dados simulados. Na Tabela 6.1, que apresenta a distribuição de frequências da categoria de respostas erradas (-1), podemos observar que a maioria dos indivíduos (34,21%) erraram somente em torno de 20% a 30% da prova. Como esse número pode ser considerado relativamente baixo, indica que os alunos estavam bem preparados para a prova. Observou-se, também, que não houve nenhum aluno que errou todas as questões da prova. O máximo de erro identificado no indivíduo 6443 que errou 78,9% da prova (Tabela 6.4).

Tabela 6.1: *Distribuição de Frequências Da Categoria De Resposta -1 (Dados Reais).*

Intervalo	Frequência Absoluta	Frequência Relativa (%)	Frequência Acumulada	Frequência Acumulada (%)
0-10%	363	5,02%	363	5,02%
10%-20%	2372	32,80%	2735	37,82%
20%-30%	2474	34,21%	5209	72,03%
30%-40%	1244	17,20%	6453	89,23%
40%-50%	666	9,21%	7119	98,44%
50%-60%	111	1,53%	7230	99,97%
60%-70%	1	0,01%	7231	99,99%
70%-80%	1	0,01%	7232	100,00%
80%-90%	0	0,00%	7232	100,00%
100%	0	0,00%	7232	100,00%
Total	7232	100%	-	-

Já na Tabela 6.2, a qual apresenta a distribuição de frequências da categoria de não respostas (0), podemos observar que a maioria dos alunos (19,29%) não respondem em torno de 30% a 40% da prova. Entretanto, em geral, 64,05% dos alunos não respondem até 40% da prova. Podemos considerar este número razoavelmente baixo, o que indica que, apesar de ser vantajoso não responder caso o aluno não tenha certeza, a maioria dos alunos deixa de responder no máximo 40% da prova, ou seja, em geral, eles respondem, ainda assim, mais da metade da prova.

A partir de uma análise mais detalhada, observou-se que 429 (5,93%) alunos respondem tudo, ou seja, não deixam nenhuma questão em branco. Dentre esses alunos há tanto alunos que acertam muito, como o indivíduo 5419, o qual acertou 78,9% da prova (Tabela 6.4), como também indivíduos que erram muito, como o indivíduo 5110, o qual acertou somente 38,4% da prova (Tabela 6.4). Dentre esses indivíduos que não deixaram nenhuma questão em branco, a maioria, 343 alunos

(79,95%), acertaram entre 50% a 70% da prova. Esse número pode ser considerado razoável, visto que o aluno que responde tudo, teoricamente, tem que ter muita certeza das respostas das questões. Considerando-se esse resultado, concluímos que estes alunos estavam respondendo as questões de forma consciente e não através do "chute".

Tabela 6.2: *Distribuição de Frequências Da Categoria De Resposta 0 (Dados Reais).*

Intervalo	Frequência Absoluta	Frequência Relativa (%)	Frequência Acumulada	Frequência Acumulada (%)
0-10%	966	13,36%	966	13,36%
10%-20%	935	12,93%	1901	26,29%
20%-30%	1336	18,47%	3237	44,76%
30%-40%	1395	19,29%	4632	64,05%
40%-50%	1188	16,43%	5820	80,48%
50%-60%	807	11,16%	6627	91,63%
60%-70%	407	5,63%	7034	97,26%
70%-80%	164	2,27%	7198	99,53%
80%-90%	31	0,43%	7229	99,96%
100%	3	0,04%	7232	100,00%
Total	7232	100%	-	-

Na Tabela 6.3, a qual apresenta a distribuição de frequências da categoria de respostas certas (1), podemos observar que poucos alunos (1,38%) acertam de 70% a 80% da prova e que somente 0,11% dos alunos acertam acima de 80% da prova. Observa-se, também, que grande parte dos alunos acerta entre 30% a 40% da prova (25,57% dos alunos) e entre 40% a 50% da prova (29,7% dos alunos). Da mesma forma, podemos verificar que um número razoável de alunos, 42,84%, acertam somente até 40% da prova. Através de uma análise mais detalhada, verificamos que não houve nenhum aluno que acertou todas as questões da prova. A maior quantidade de questões corretas foi obtida pelo indivíduo 4711, que acertou 86,5% das questões da prova (Tabela 6.4).

Tabela 6.3: *Distribuição de Frequências Da Categoria De Resposta 1 (Dados Reais).*

Intervalo	Frequência Absoluta	Frequência Relativa (%)	Frequência Acumulada	Frequência Acumulada (%)
0-10%	16	0,22%	16	0,22%
10%-20%	238	3,29%	254	3,51%
20%-30%	995	13,76%	1249	17,27%
30%-40%	1849	25,57%	3098	42,84%
40%-50%	2148	29,70%	5246	72,54%
50%-60%	1447	20,01%	6693	92,55%
60%-70%	431	5,96%	7124	98,51%
70%-80%	100	1,38%	7224	99,89%
80%-90%	8	0,11%	7232	100,00%
100%	0	0,00%	7232	100,00%
Total	7232	100%	-	-

Na Tabela 6.4 são apresentadas as frequências das categorias de respostas de alguns indivíduos. Da mesma forma como analisado nos dados simulados apresentados no capítulo anterior, podemos observar diferentes perfis de respondentes. O indivíduo 3 representa um indivíduo regular, que tem um conhecimento mediano do assunto e responde as questões com um pouco de cautela. Esse perfil pode ser identificado pelo fato deste indivíduo ter acertado, errado e não respondido aproximadamente a mesma quantidade de questões.

O indivíduo 4 pode representar um indivíduo que tem um razoável conhecimento do assunto abordado e é muito cauteloso, pois ele deixou de responder muitas questões, acertou poucas e não errou quase nada. Da mesma forma, o indivíduo 5311 pode ter as mesmas características que o indivíduo anterior, porém tem um domínio do assunto abordado ainda mais baixo, acertando menos questões e deixando mais questões em branco que o indivíduo 4. Ainda no perfil de indivíduos cautelosos ou que não respondem muitas questões da prova, podemos citar o indivíduo 6825, o qual foi o respondente com maior índice de questões em branco, deixando de responder 94% da prova, acertando pouquíssimas questões, mas ainda assim errou (3,8%), e acertou somente (2,11%).

Em relação ao indivíduo 4035, podemos concluir que ele é um respondente extremamente cauteloso, porém com alto domínio do assunto, pois das 115 questões que respondeu, acertou 108, errando somente 7 questões. Apesar de ter acertado muitas questões, ele ainda deixou 122 questões em branco. Dessa forma, este indivíduo representa um indivíduo que faz a prova de forma consciente, de acordo com o que

sabe.

Conforme já falado anteriormente, o indivíduo 5110 representa um indivíduo que não teve cautela alguma ao responder a prova, respondendo todas as questões, e mesmo sem ter um bom domínio do assunto da prova, pois errou 61,6% do exame. Esse indivíduo representa o respondente que mesmo não tendo certeza da resposta correta da questão, "chuta". Por outro lado, como também já citado na análise anterior, o indivíduo 5419 também é pouco cauteloso, podendo ser também um indivíduo que "chuta", pois respondeu todas as questões, entretanto, obteve um alto índice de acerto (187 questões), representando assim um indivíduo que tem alto conhecimento do assunto medido.

Ainda na mesma linha de raciocínio, podemos citar também o indivíduo 4711, o qual foi aquele que obteve o maior percentual de acerto da prova, sendo este um respondente com alto domínio do conteúdo, mas que também podendo ser caracterizado como um indivíduo não cauteloso, pois deixou somente 2 questões em branco e errou 30 questões. E por último, neste sentido, podemos citar o indivíduo 37, que respondeu todas as questões, mas acertou praticamente a mesma quantidade de questões que errou, obtendo, por exemplo, pontuação praticamente 0 no método de correção convencional, o qual será abordado com mais detalhes posteriormente.

Para o perfil de um respondente pouco cauteloso e que também não sabe muito do assunto, podemos citar o indivíduo 6443, o qual também foi o candidato que obteve o maior percentual de erros da prova (78,9% das questões). Este indivíduo respondeu praticamente todas as questões da prova e acertou somente 20,25% da prova. Dessa forma, podemos concluir que há diferentes perfis de indivíduos que responderam a prova e, portanto, os dados simulados, nesse ponto, representaram bem a situação real.

Tabela 6.4: *Frequências de respostas por indivíduo (Dados Reais).*

	Respostas			Percentual de Respostas		
Indivíduos	-1	0	1	-1	0	1
3	62	87	88	26,16%	36,71%	37,13%
4	7	175	55	2,95%	73,84%	23,21%
6	66	81	90	27,85%	34,18%	37,97%
7	48	74	115	20,25%	31,22%	48,52%
8	70	38	129	29,54%	16,03%	54,43%
9	24	161	52	10,13%	67,93%	21,94%
19	32	56	149	13,50%	23,63%	62,87%
37	116	0	121	48,95%	0,00%	51,05%
4035	7	122	108	2,95%	51,48%	45,57%
4711	30	2	205	12,66%	0,84%	86,50%
5110	146	0	91	61,60%	0,00%	38,40%
5311	3	199	35	1,27%	83,97%	14,77%
5419	50	0	187	21,10%	0,00%	78,90%
6443	187	2	48	78,90%	0,84%	20,25%
6825	9	223	5	3,80%	94,09%	2,11%

Na tabela 6.5, são apresentadas as frequências de respostas de cada categoria de forma global. Num total de 1.713.984 respostas - correspondente às respostas de 7232 indivíduos às 237 questões da prova -, foi observado que 24,81% das questões foram marcadas de forma incorreta, 32,88% das questões foram deixadas em branco e 42,23% das questões foram marcadas corretamente.

Tabela 6.5: *Frequências Globais de Respostas (Dados Reais).*

	Respostas		
	-1	0	1
Frequência	425241	563620	725123
Percentual	24,81%	32,88%	42,23%

Podemos observar que as quantidades de respostas corretas (1), incorretas (-1) e ausentes (0) ficaram relativamente próximas, indicando um equilíbrio, de forma geral, entre a distribuição das categorias. Comparando com esta mesma análise referente aos dados simulados, podemos observar uma diferença na distribuição de quantidade de categorias, em que nos dados simulados obtivemos um percentual de não respostas um pouco maior. Assim, pode-se concluir que a característica de penalização ao se errar uma questão da prova de fato faz com que os indivíduos não respondam muitas questões, entretanto, não tanto quanto foi representando pelos dados simulados.

Por outro lado, ao analisarmos o percentual das quantidades de respostas

corretas e incorretas nos dados reais, podemos observar praticamente a mesma diferença que obtivemos entre essas mesmas quantidades nos dados simulados, sendo estas 17,42% e 19,17% para os dados reais e para os dados simulados, respectivamente. Essa distribuição de frequência global entre as categorias está representada graficamente na Figura 6.1.

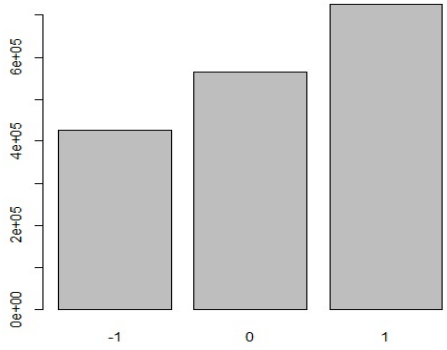

Figura 6.1: Histograma das frequências de respostas (Dados Reais).

6.2 Análise dos Parâmetros dos Itens e Curva Característica dos Itens

Após a construção do banco de dados, como já dito anteriormente, foi aplicado o modelo politômico da Teoria de Resposta ao Item, cujo nome é Modelo de Resposta Gradual (MRG), com intuito de se estimar as proficiências dos indivíduos considerando-se o método proposto. A partir da aplicação do modelo, foram estimados os parâmetros dos itens e geradas as respectivas curvas característica dos itens.

Na Tabela 6.6, são apresentadas algumas medidas descritivas dos parâmetros dos itens estimados a partir do MRG. A partir disso, podemos observar que o valor mínimo para o parâmetro de discriminação foi -0,639, o que não é esperado, já que a probabilidade de acerto deve crescer a medida que a profiência aumenta, ou seja, o valor do parametro de discriminação deve ser sempre maior que 0. Entretanto, podemos afirmar também que houve itens que discriminaram bem os indivíduos, pois o valor máximo para o parâmetro de discriminação resultou em 1,595, sendo este

maior do que o ponto de corte indicado pela literatura, valor 1.

Valores negativos de a_i não são esperados, pois a presença destes valores faz com que a probabilidade de acerto do item diminua com o aumento da habilidade sobre o assunto avaliado. De acordo com Back (2001), valores negativos do parâmetro de discriminação indicam que há algo errado com o item, podendo estar mal escrito ou havendo alguma desinformação prevalente entre os alunos de alta capacidade.

Com relação aos parâmetros de dificuldade dos itens (b_{i1} e b_{i2}), podemos observar que há itens com categorias com baixa dificuldade, com o mínimo de -3000, e com alta dificuldade, com máximo de 637. Esses valores resultaram completamente fora do intervalo em que a literatura indica (-2 a +2). Entretanto, em média, os valores desses parâmetros estão mais próximos do intervalo recomendado.

Baseado na idenficação de valores não ideais tanto para os parâmetros de discriminação, como para os de dificuldade, podemos supor que há itens na análise que não são informativos, ou seja, que não avaliam bem os indivíduos. Talvez esses valores levantem indícios de que alguns itens devessem ser retirados da prova antes de ser aplicada. Entretanto, como este tipo de análise não é o foco deste estudo, nenhum item foi excluído da análise, pois a prova já foi aplicada.

Tabela 6.6: *Análise Descritiva dos parâmetros dos itens (Modelo de Resposta Gradual-Dados Reais).*

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Parâmetros de Discriminação	-0,639	0,120	0,378	0,396	0,646	1,595
Parâmetros de Dificuldade (b_{i1})	-3.000,000	-5,034	-2,508	-13,770	-1,179	375,200
Parâmetros de Dificuldade (b_{i2})	-45,280	-1,447	0,207	4,105	2,038	637,400

Nas Tabelas 6.7, 6.8 e 6.9, são apresentados os valores dos parâmetros dos itens estimados e os respectivos percentuais de respostas de cada categoria. Analisando o item 1, podemos observar que esse item pode ser considerado um item difícil, pois contém os dois parâmetros de dificuldade ($b_{1,1}$ e $b_{1,2}$) positivos e relativamente altos. Observa-se também que o número de respostas erradas condiz com os altos valores dos parâmetros de dificuldade, pois 56,24% dos indivíduos erraram a questão. Além disso, podemos perceber que o número de não respostas é maior que o número de acertos, confirmando a dificuldade do item. Em relação ao índice de discriminação, podemos afirmar que este item não discrimina tão bem os indivíduos.

Tabela 6.7: *Frequência das respostas por Item e Parâmetros dos itens 1 a 80 (Dados Reais).*

	Categoria			Parâmetros					Categoria			Parâmetros			
Itens	-1 (%)	0 (%)	1 (%)	a_i	b_{i1}	b_{i2}	$b_{i,\text{médio}}$	Itens	-1 (%)	0 (%)	1 (%)	a_i	b_{i1}	b_{i2}	$b_{i,\text{médio}}$
Item 1	56,24	15,89	27,88	0,407	0,595	2,380	1,488	Item 41	13,88	18,35	67,77	0,680	-2,895	-1,214	-2,055
Item 2	60,98	17,55	21,47	-0,097	-4,643	-13,436	-9,040	Item 42	16,40	16,87	66,73	0,363	-4,629	-2,010	-3,320
Item 3	27,02	18,24	54,74	0,156	-6,417	-1,251	-3,834	Item 43	10,16	32,15	57,69	0,809	-2,954	-0,474	-1,714
Item 4	45,28	30,13	24,59	0,205	-0,974	5,484	2,255	Item 44	31,04	43,07	25,88	0,150	-5,413	7,012	0,800
Item 5	23,20	17,60	59,20	0,806	-1,709	-0,584	-1,147	Item 45	15,67	29,08	55,25	0,176	-9,625	-1,233	-5,429
Item 6	54,60	16,51	28,89	-0,127	-1,458	-7,126	-4,292	Item 46	11,50	46,36	42,13	0,534	-3,995	0,612	-1,692
Item 7	47,32	32,08	20,60	0,100	-1,088	13,523	6,218	Item 47	9,98	29,07	60,95	0,090	-24,547	-4,976	-14,762
Item 8	58,05	17,87	24,09	-0,235	-1,426	-4,971	-3,199	Item 48	19,94	45,98	34,08	0,356	-3,990	1,911	-1,040
Item 9	10,34	9,75	79,91	0,515	-4,383	-2,834	-3,609	Item 49	43,79	11,26	44,95	0,847	-0,404	0,207	-0,099
Item 10	14,80	24,99	60,22	0,674	-2,825	-0,724	-1,775	Item 50	30,86	42,64	26,49	0,075	-10,769	13,554	1,393
Item 11	7,98	12,62	79,40	0,701	-3,745	-2,122	-2,934	Item 51	13,08	23,89	63,03	1,374	-1,743	-0,568	-1,156
Item 12	40,83	20,69	38,48	0,293	-1,301	1,623	0,161	Item 52	14,75	68,07	17,17	-0,192	9,199	-8,220	0,490
Item 13	19,88	29,19	50,93	0,477	-3,087	-0,138	-1,613	Item 53	19,34	35,69	44,97	0,497	-3,020	0,403	-1,309
Item 14	25,08	24,47	50,44	0,599	-1,987	-0,076	-1,032	Item 54	36,60	30,39	33,01	0,567	-1,053	1,315	0,131
Item 15	25,07	22,98	51,95	0,646	-1,849	-0,162	-1,006	Item 55	38,91	25,57	35,52	0,335	-1,393	1,810	0,209
Item 16	65,45	14,02	20,53	0,142	4,498	9,564	7,031	Item 56	35,62	34,91	29,47	0,538	-1,242	1,647	0,203
Item 17	41,62	32,70	25,68	-0,138	2,411	-7,798	-2,694	Item 57	13,81	59,08	27,10	-0,130	14,192	-7,650	3,271
Item 18	26,19	29,98	43,83	0,464	-2,384	0,498	-0,943	Item 58	20,59	26,95	52,46	-0,006	222,627	16,219	119,423
Item 19	67,81	14,45	17,74	0,334	2,276	4,692	3,484	Item 59	11,50	28,68	59,82	0,633	-3,445	-0,721	-2,083
Item 20	15,40	46,92	37,68	0,156	-10,963	3,247	-3,858	Item 60	13,88	20,81	65,31	0,460	-4,128	-1,463	-2,796
Item 21	21,20	22,76	56,04	0,718	-2,014	-0,402	-1,208	Item 61	10,99	27,17	61,84	0,899	-2,612	-0,663	-1,638
Item 22	39,68	25,18	35,14	0,687	-0,722	0,927	0,103	Item 62	17,23	16,34	66,43	-0,004	375,234	163,139	269,187
Item 23	11,06	13,32	75,62	1,093	-2,228	-1,268	-1,748	Item 63	8,25	24,06	67,69	0,472	-5,269	-1,650	-3,460
Item 24	20,74	29,01	50,25	0,030	-45,263	-0,370	-22,817	Item 64	11,92	24,92	63,16	0,781	-2,807	-0,807	-1,807
Item 25	44,98	14,26	40,76	0,455	-0,477	0,847	0,185	Item 65	15,29	13,11	71,60	0,696	-2,682	-1,488	-2,085
Item 26	11,26	37,62	51,12	0,381	-5,540	-0,120	-2,830	Item 66	29,12	35,77	35,11	0,309	-2,953	2,016	-0,469
Item 27	9,29	32,22	58,49	0,602	-3,988	-0,633	-2,311	Item 67	6,42	17,44	76,15	0,793	-3,657	-1,651	-2,654
Item 28	7,76	14,55	77,70	0,539	-4,803	-2,461	-3,632	Item 68	30,66	40,64	28,71	0,327	-2,592	2,814	0,111
Item 29	18,13	19,80	62,07	1,023	-1,742	-0,618	-1,180	Item 69	14,49	14,28	71,23	0,360	-5,077	-2,624	-3,851
Item 30	36,57	15,50	47,93	0,346	-1,647	0,235	-0,706	Item 70	49,57	23,71	26,71	0,325	-0,129	3,107	1,489
Item 31	20,78	16,27	62,94	0,871	-1,744	-0,717	-1,231	Item 71	6,66	15,11	78,22	1,036	-2,889	-1,485	-2,187
Item 32	79,83	8,46	11,71	-0,639	-2,334	-3,381	-2,858	Item 72	8,96	24,63	66,41	0,753	-3,331	-1,036	-2,184
Item 33	8,10	9,03	82,87	1,595	-1,967	-1,354	-1,661	Item 73	23,84	25,72	50,44	0,713	-1,825	-0,087	-0,956
Item 34	22,70	25,79	51,51	0,715	-1,918	-0,151	-1,035	Item 74	22,50	38,86	38,65	0,353	-3,618	1,310	-1,154
Item 35	37,22	18,63	44,15	0,214	-2,508	1,075	-0,717	Item 75	7,52	16,41	76,06	0,307	-8,309	-3,859	-6,084
Item 36	16,32	14,49	69,19	0,452	-3,764	-1,885	-2,825	Item 76	17,06	28,25	54,69	0,718	-2,425	-0,344	-1,385
Item 37	27,97	8,74	63,29	0,966	-1,179	-0,707	-0,943	Item 77	37,82	17,48	44,70	0,706	-0,830	0,276	-0,277
Item 38	13,86	39,30	46,85	0,701	-2,824	0,162	-1,331	Item 78	29,30	18,85	51,85	1,214	-0,978	-0,167	-0,573
Item 39	40,27	28,03	31,71	0,447	-0,961	1,751	0,395	Item 79	21,23	13,87	64,91	0,151	-8,757	-4,126	-6,442
Item 40	4,23	10,26	85,51	0,885	-3,831	-2,255	-3,043	Item 80	15,63	14,13	70,24	0,949	-2,065	-1,110	-1,588

Tabela 6.8: *Frequência das respostas por Item e Parâmetros dos itens 81 a 160 (Dados Reais).*

	Categoria			Parâmetros					Categoria			Parâmetros			
Itens	-1 (%)	0 (%)	1 (%)	a_i	b_{i1}	b_{i2}	$b_{i,\text{médio}}$	Itens	-1 (%)	0 (%)	1 (%)	a_i	b_{i1}	b_{i2}	$b_{i,\text{médio}}$
Item 81	10,81	21,27	67,92	0,613	-3,648	-1,337	-2,493	Item 121	27,86	22,93	49,21	0,268	-3,620	0,098	-1,761
Item 82	5,42	8,52	86,06	0,583	-5,117	-3,302	-4,210	Item 122	29,94	16,04	54,02	0,095	-8,968	-1,696	-5,332
Item 83	22,62	22,22	55,16	0,764	-1,794	-0,330	-1,062	Item 123	9,82	38,80	51,38	0,296	-7,569	-0,166	-3,868
Item 84	4,76	15,64	79,60	0,541	-5,734	-2,651	-4,193	Item 124	32,29	35,44	32,27	0,333	-2,275	2,280	0,002
Item 85	15,97	42,04	41,99	0,524	-3,337	0,622	-1,358	Item 125	33,02	46,31	20,67	0,140	-5,056	9,634	2,289
Item 86	25,51	23,48	51,01	1,089	-1,217	-0,098	-0,658	Item 126	15,60	35,97	48,44	0,385	-4,497	0,166	-2,166
Item 87	46,46	27,74	25,80	0,375	-0,396	2,901	1,253	Item 127	17,95	19,62	62,43	0,142	-10,755	-3,595	-7,175
Item 88	10,02	20,99	68,99	0,603	-3,849	-1,447	-2,648	Item 128	55,57	34,00	10,43	-0,168	-1,336	-12,832	-7,084
Item 89	13,58	18,27	68,16	0,153	-12,196	-5,034	-8,615	Item 129	37,57	46,71	15,72	-0,118	4,295	-14,222	-4,964
Item 90	28,08	28,72	43,20	0,649	-1,609	0,410	-0,600	Item 130	8,28	40,98	50,73	0,342	-7,141	-0,079	-3,610
Item 91	6,73	15,83	77,43	0,496	-5,478	-2,610	-4,044	Item 131	22,23	30,19	47,58	-0,033	37,455	-2,926	17,265
Item 92	46,09	37,80	16,11	-0,175	0,862	-9,500	-4,319	Item 132	37,89	27,43	34,68	-0,468	1,061	-1,461	-0,200
Item 93	16,81	25,79	57,40	0,998	-1,882	-0,412	-1,147	Item 133	31,29	27,27	41,44	0,323	-2,519	1,068	-0,726
Item 94	31,08	35,15	33,77	0,793	-1,153	0,927	-0,113	Item 134	17,27	14,96	67,77	0,630	-2,661	-1,289	-1,975
Item 95	32,95	46,10	20,95	-0,211	3,374	-6,370	-1,498	Item 135	45,91	28,18	25,91	0,030	-5,530	35,107	14,789
Item 96	15,92	39,28	44,80	0,239	-7,054	0,856	-3,099	Item 136	34,39	31,37	34,24	0,582	-1,224	1,166	-0,029
Item 97	15,20	56,21	28,60	0,268	-6,483	3,463	-1,510	Item 137	25,90	39,81	34,29	0,555	-2,037	1,205	-0,416
Item 98	47,39	20,95	31,66	0,341	-0,343	2,287	0,972	Item 138	11,85	42,91	45,24	0,396	-5,191	0,494	-2,349
Item 99	31,31	26,84	41,86	0,032	-24,463	10,198	-7,133	Item 139	16,33	43,69	39,98	0,095	-17,137	4,284	-6,427
Item 100	41,51	25,47	33,02	0,598	-0,671	1,219	0,274	Item 140	25,58	27,65	46,76	-0,324	3,343	-0,443	1,450
Item 101	31,44	13,33	55,23	0,768	-1,161	-0,342	-0,752	Item 141	17,53	53,39	29,08	-0,038	40,323	-23,142	8,591
Item 102	9,07	37,43	53,50	0,047	-49,516	-2,971	-26,244	Item 142	27,61	52,42	19,97	-0,083	11,677	-16,838	-2,581
Item 103	8,52	57,11	34,38	0,461	-5,287	1,492	-1,898	Item 143	30,17	49,45	20,38	0,320	-2,669	4,364	0,848
Item 104	3,77	11,12	85,11	0,841	-4,150	-2,314	-3,232	Item 144	24,96	46,92	28,13	0,500	-2,263	2,023	-0,120
Item 105	5,56	10,65	83,79	0,308	-9,305	-5,422	-7,364	Item 145	46,67	17,80	35,54	0,661	-0,271	0,938	0,334
Item 106	21,28	45,45	33,27	0,597	-2,316	1,271	-0,523	Item 146	11,45	65,60	22,95	0,506	-4,198	2,531	-0,834
Item 107	3,80	6,11	90,09	0,172	-18,817	-12,868	-15,843	Item 147	19,93	57,29	22,79	-0,110	12,713	-11,114	0,799
Item 108	12,13	18,81	69,07	0,318	-6,337	-2,581	-4,459	Item 148	24,61	40,64	34,75	0,311	-3,700	2,038	-0,831
Item 109	3,18	8,88	87,94	0,925	-4,021	-2,426	-3,224	Item 149	20,34	43,58	36,08	0,686	-2,196	0,849	-0,674
Item 110	35,37	14,91	49,72	0,804	-0,895	-0,040	-0,468	Item 150	45,98	15,85	38,18	0,605	-0,327	0,822	0,248
Item 111	45,95	31,97	22,08	-0,263	0,598	-4,882	-2,142	Item 151	17,44	52,79	29,77	0,910	-1,961	1,039	-0,461
Item 112	17,37	51,31	31,32	0,886	-2,023	0,945	-0,539	Item 152	41,19	33,95	24,86	0,345	-1,054	3,293	1,120
Item 113	17,24	12,03	70,73	0,939	-1,922	-1,117	-1,520	Item 153	47,82	16,14	36,05	0,075	-1,175	7,651	3,238
Item 114	18,21	36,39	45,40	0,082	-18,320	2,248	-8,036	Item 154	23,42	46,36	30,21	0,092	-12,889	9,095	-1,897
Item 115	62,51	18,68	18,81	-0,318	-1,656	-4,695	-3,176	Item 155	10,38	73,31	16,30	-0,132	16,487	-12,427	2,030
Item 116	52,93	21,86	25,21	0,176	0,652	6,216	3,434	Item 156	21,54	47,48	30,97	0,474	-2,849	1,760	-0,545
Item 117	31,44	26,81	41,75	0,122	-6,450	2,713	-1,869	Item 157	26,83	48,56	24,61	0,084	-12,016	13,360	0,672
Item 118	30,67	30,01	39,33	0,179	-4,564	2,456	-1,054	Item 158	17,98	56,58	25,44	0,418	-3,733	2,677	-0,528
Item 119	37,43	29,94	32,63	0,037	-14,059	19,841	2,891	Item 159	11,23	49,59	39,19	0,350	-6,023	1,275	-2,374
Item 120	37,51	17,51	44,98	1,015	-0,651	0,182	-0,235	Item 160	14,96						

Tabela 6.9: *Frequência das respostas por Item e Parâmetros dos itens 161 a 237 (Dados Reais).*

	Categoria			Parâmetros					Categoria			Parâmetros			
Itens	-1 (%)	0 (%)	1 (%)	a_i	b_{i1}	b_{i2}	$b_{i,\text{médio}}$	Itens	-1 (%)	0 (%)	1 (%)	a_i	b_{i1}	b_{i2}	$b_{i,\text{médio}}$
Item 161	24,74	33,90	41,36	0,223	-5,034	1,575	-1,730	Item 200	15,29	47,90	36,81	0,936	-2,094	0,622	-0,736
Item 162	27,81	36,08	36,12	0,096	-9,894	5,952	-1,971	Item 201	15,65	70,55	13,80	-0,275	6,193	-6,757	-0,282
Item 163	49,06	21,61	29,33	0,187	-0,224	4,721	2,249	Item 202	31,10	42,45	26,45	0,559	-1,548	1,908	0,180
Item 164	14,98	32,45	52,57	0,069	-25,215	-1,521	-13,368	Item 203	14,01	31,97	54,02	0,531	-3,575	-0,331	-1,953
Item 165	19,21	31,37	49,42	0,545	-2,792	0,017	-1,388	Item 204	20,49	41,30	38,21	0,982	-1,651	0,500	-0,576
Item 166	10,79	77,18	12,03	0,020	-105,846	99,746	-3,050	Item 205	66,76	19,88	13,36	-0,041	-16,921	-45,281	-31,101
Item 167	11,09	74,96	13,95	0,082	-25,505	22,385	-1,560	Item 206	23,53	25,58	50,88	0,511	-2,437	-0,101	-1,269
Item 168	32,25	43,85	23,91	0,179	-4,179	6,484	1,153	Item 207	44,10	32,63	23,27	0,148	-1,675	8,028	3,177
Item 169	22,80	34,32	42,88	0,574	-2,279	0,499	-0,890	Item 208	23,24	35,55	41,21	0,533	-2,397	0,665	-0,866
Item 170	16,98	51,62	31,40	0,286	-5,629	2,796	-1,417	Item 209	34,18	33,06	32,76	0,463	-1,530	1,569	0,020
Item 171	31,75	45,58	22,68	0,073	-10,521	16,812	3,146	Item 210	11,02	53,89	35,09	0,837	-2,756	0,804	-0,976
Item 172	12,26	68,75	18,99	-0,215	9,242	-6,793	1,225	Item 211	12,47	65,69	21,83	0,378	-5,269	3,468	-0,901
Item 173	20,93	63,51	15,56	0,051	-26,145	33,287	3,571	Item 212	15,92	54,30	29,78	0,313	-5,412	2,784	-1,314
Item 174	15,04	36,23	48,73	0,637	-2,901	0,070	-1,416	Item 213	39,33	19,14	41,54	0,585	-0,816	0,607	-0,105
Item 175	16,94	45,17	37,89	0,385	-4,260	1,296	-1,482	Item 214	23,37	38,52	38,11	0,819	-1,650	0,622	-0,514
Item 176	20,93	39,64	39,42	0,120	-11,178	3,549	-3,815	Item 215	34,71	20,75	44,54	0,812	-0,922	0,257	-0,333
Item 177	46,53	14,64	38,83	1,069	-0,254	0,422	0,084	Item 216	26,09	50,36	23,55	-0,128	8,119	-9,304	-0,593
Item 178	23,42	17,34	59,24	-0,147	8,076	2,535	5,306	Item 217	12,38	56,17	31,46	0,410	-4,903	1,961	-1,471
Item 179	20,78	58,86	20,35	-0,042	32,250	-32,875	-0,313	Item 218	27,65	42,16	30,19	0,376	-2,674	2,241	-0,217
Item 180	35,25	17,57	47,18	0,637	-1,092	0,129	-0,482	Item 219	30,06	43,45	26,49	0,243	-3,524	4,235	0,356
Item 181	11,41	34,40	54,19	1,019	-2,320	-0,259	-1,290	Item 220	15,09	69,30	15,61	0,029	-59,052	57,684	-0,684
Item 182	20,92	68,57	10,51	-0,065	20,589	-33,141	-6,276	Item 221	27,46	45,20	27,34	0,658	-1,618	1,601	-0,009
Item 183	19,70	45,67	34,62	0,111	-12,791	5,707	-3,542	Item 222	23,71	37,74	38,55	0,612	-2,079	0,773	-0,653
Item 184	23,91	48,35	27,74	0,371	-3,279	2,576	-0,352	Item 223	25,65	44,32	30,03	0,736	-1,617	1,238	-0,190
Item 185	23,59	39,48	36,93	0,295	-4,098	1,797	-1,151	Item 224	20,11	59,96	19,94	-0,066	21,053	-21,223	-0,085
Item 186	19,77	8,10	72,12	0,538	-2,761	-1,892	-2,327	Item 225	19,28	56,89	23,84	-0,198	7,284	-5,937	0,674
Item 187	25,82	48,81	25,37	0,158	-6,755	6,840	0,043	Item 226	25,33	39,89	34,78	0,462	-2,421	1,438	-0,492
Item 188	23,17	29,58	47,25	0,692	-1,952	0,087	-0,933	Item 227	34,15	38,66	27,18	0,064	-10,148	15,357	2,605
Item 189	26,20	27,21	46,58	0,917	-1,330	0,123	-0,604	Item 228	21,49	45,52	32,99	0,232	-5,611	3,146	-1,233
Item 190	24,56	34,64	40,80	0,537	-2,219	0,713	-0,753	Item 229	25,71	46,94	27,35	0,025	-42,404	39,110	-1,647
Item 191	51,24	18,81	29,95	0,580	0,010	1,493	0,752	Item 230	22,59	49,71	27,70	-0,155	8,021	-6,168	0,927
Item 192	30,72	47,72	21,56	0,245	-3,355	5,352	0,999	Item 231	25,62	23,11	51,27	0,760	-1,581	-0,119	-0,850
Item 193	17,49	26,27	56,24	1,026	-1,781	-0,342	-1,062	Item 232	18,54	30,85	50,61	0,826	-2,021	-0,080	-1,051
Item 194	18,99	33,09	47,93	0,609	-2,553	0,116	-1,219	Item 233	15,47	27,21	57,31	0,580	-3,104	-0,569	-1,837
Item 195	38,03	16,68	45,30	1,036	-0,655	0,121	-0,267	Item 234	14,42	44,93	40,65	0,001	-2999,520	637,359	-1181,081
Item 196	19,61	21,10	59,29	1,054	-1,629	-0,508	-1,069	Item 235	34,67	42,33	23,01	0,416	-1,583	3,010	0,714
Item 197	17,10	72,41	10,48	-0,063	24,949	-33,883	-4,467	Item 236	23,51	48,31	28,18	0,211	-5,616	4,487	-0,565
Item 198	18,06	39,64	42,30	0,167	-9,171	1,830	-3,671	Item 237	17,35	51,62	31,03	-0,150	10,517	-5,310	2,604
Item 199	21,58	41,30	37,11	0,534	-2,552	1,026	-0,763								

Analisando o item 215, podemos concluir que este pode representar um item não muito fácil e não muito difícil, pois o valor do parâmetro $b_{215,2}$ é positivo, porém baixo. As quantidades de respostas entre as categorias ficaram parcialmente distribuídas entre as três opções, sendo que a categoria de resposta 0, obteve o maior valor (44,53% das respostas).

Já para o item 210, podemos perceber indícios de que seja um item fácil por conter um elevado percentual de respostas corretas e possuir baixos valores dos parâmetros de dificuldade. Em relação ao parâmetro de discriminação, podemos perceber uma discriminação relativa entre os indivíduos, pois o valor de a_{210} está bem próximo de 1.

Analisando o item 113, podemos perceber um alto percentual de não respostas, porém os parâmetros de dificuldade resultaram em baixos valores. Podemos supor, por exemplo, que este item possa ter abordado um assunto não muito esperado pelos alunos ou também um assunto não muito discutido para o vestibular. O índice de discriminação resultou um valor muito próximo do ideal.

Após a análise do valor de cada parâmetro dos itens, foram construídas as as respectivas Curvas Característica dos Itens (CCI), as quais representam a relação existente entre a probabilidade de o indivíduo acertar o item dado a sua habilidade, $P(U_{ij} = 1|\theta_j)$, e os parâmetros do modelo, como já mencionado. Da mesma forma que as análises do capítulo anterior, nos gráficos de CCI o eixo horizontal representa o valor do traço latente medido e o eixo vertical corresponde à probabilidade de o indivíduo ter sua resposta classificada em uma das três categorias (-1, 0, 1).

A título de exemplificação, a Figura 6.2 ilustra a CCI do item 112. Neste gráfico, podemos observar que indivíduos com habilidade até aproximadamente -2 têm mais probabilidade de errar a questão, "escolhendo", portanto, a categoria -1. Já indivíduos com habilidade aproximadamente de -2 a 0,9 têm maior probabilidade de não responder a questão. E, por fim, indivíduos com habilidade acima de 0,9, aproximadamente, têm maior chance de acertar a questão. Relacionando este gráfico com os valores da Tabela 6.8, podemos perceber que este item tem um bom índice de discriminação, com a_{112}=0,886, o que condiz com o bom espaçamento entre as curvas, confirmando uma boa discriminação.

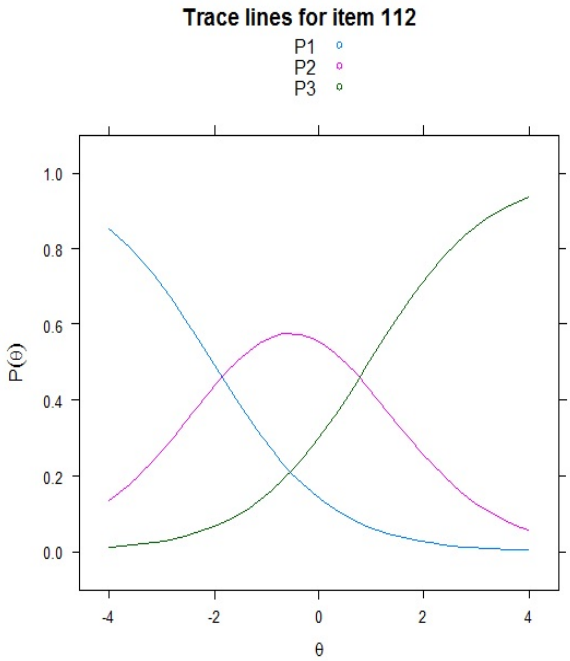

Figura 6.2: Curva Característica do Item 112 (Dados Reais).

Na Figura 6.3, estão apresentadas as CCIs dos itens 38, 151 e 94. Comparando os itens 38 e 151, podemos perceber que o item 151 é mais difícil que o item 38. Isso se dá pelo fato de que um indivíduo para acertar o item 151 precisa de uma habilidade acima de 1, aproximadamente, e já no item 38, para obter este mesmo resultado, o indivíduo precisa ter uma habilidade acima de somente -0,9, aproximadamente. Comparando agora os itens 151 e 94, é possível perceber que o item 151 discrimina melhor os indivíduos que o item 94, pois, neste último, além do valor do parâmetro de discriminação ser menor (Tabelas 6.7 e 6.8), podemos observar uma proximidade das curvas das categorias do item.

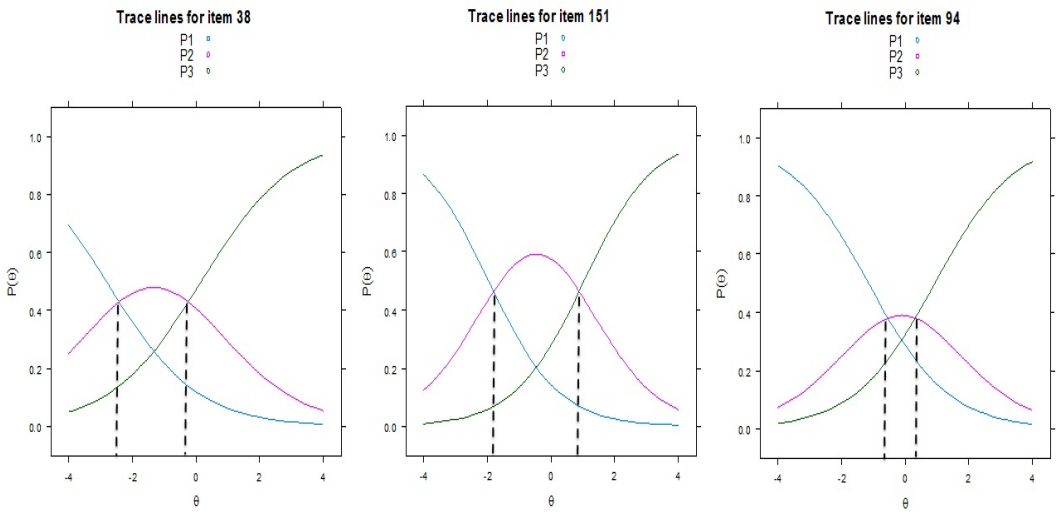

Figura 6.3: Curvas Característica dos Itens 38, 151 e 94. (Dados Reais).

Da mesma forma como verificado nas análises dos dados simulados, percebemos itens tanto com valores extremos e fora do intervalo esperado em relação aos parâmetros dos itens, discriminação e dificuldade, como também com curvas das categorias que não se sobressaíram em nenhuma região, o que indica falta de qualidade para avaliar o traço latente. Podemos citar, por exemplo, os itens 234, 62, 166, 220, entre outros.

Do mesmo modo como já citado no início desta seção e também no capítulo anterior, itens com este tipo de comportamento representam itens não informativos e que não avaliam bem os indivíduos. Entretanto, a análise não foi refeita retirando-se esses itens pelo fato de essa análise não ser o objetivo principal deste trabalho, visto que, como também já mencionado, a prova já foi aplicada. Nas Figuras 6.4 a 6.8, são apresentadas as CCI de todos os itens analisados.

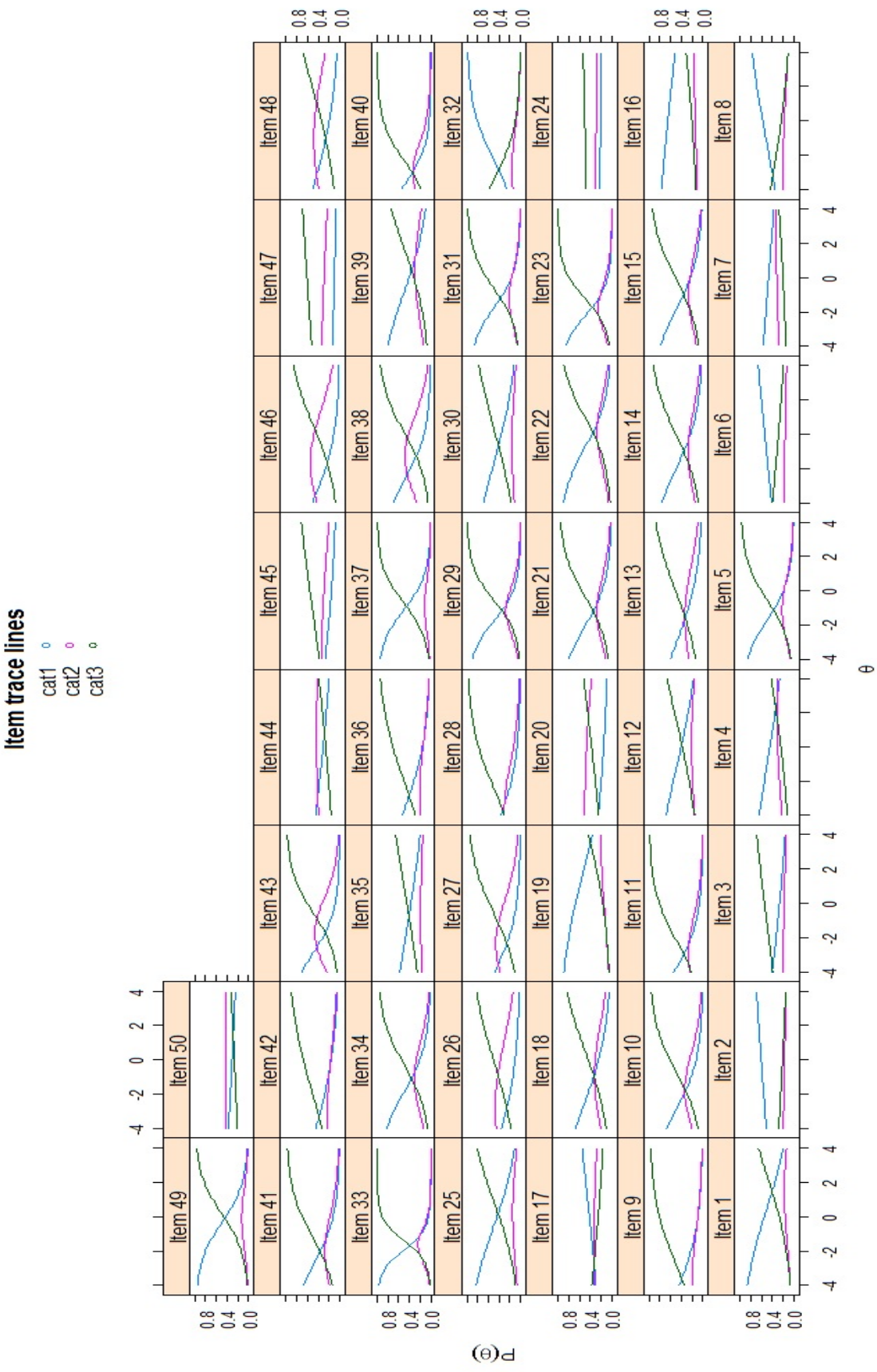

Figura 6.4: Curvas Características dos Itens 1 a 50 (Dados Reais).

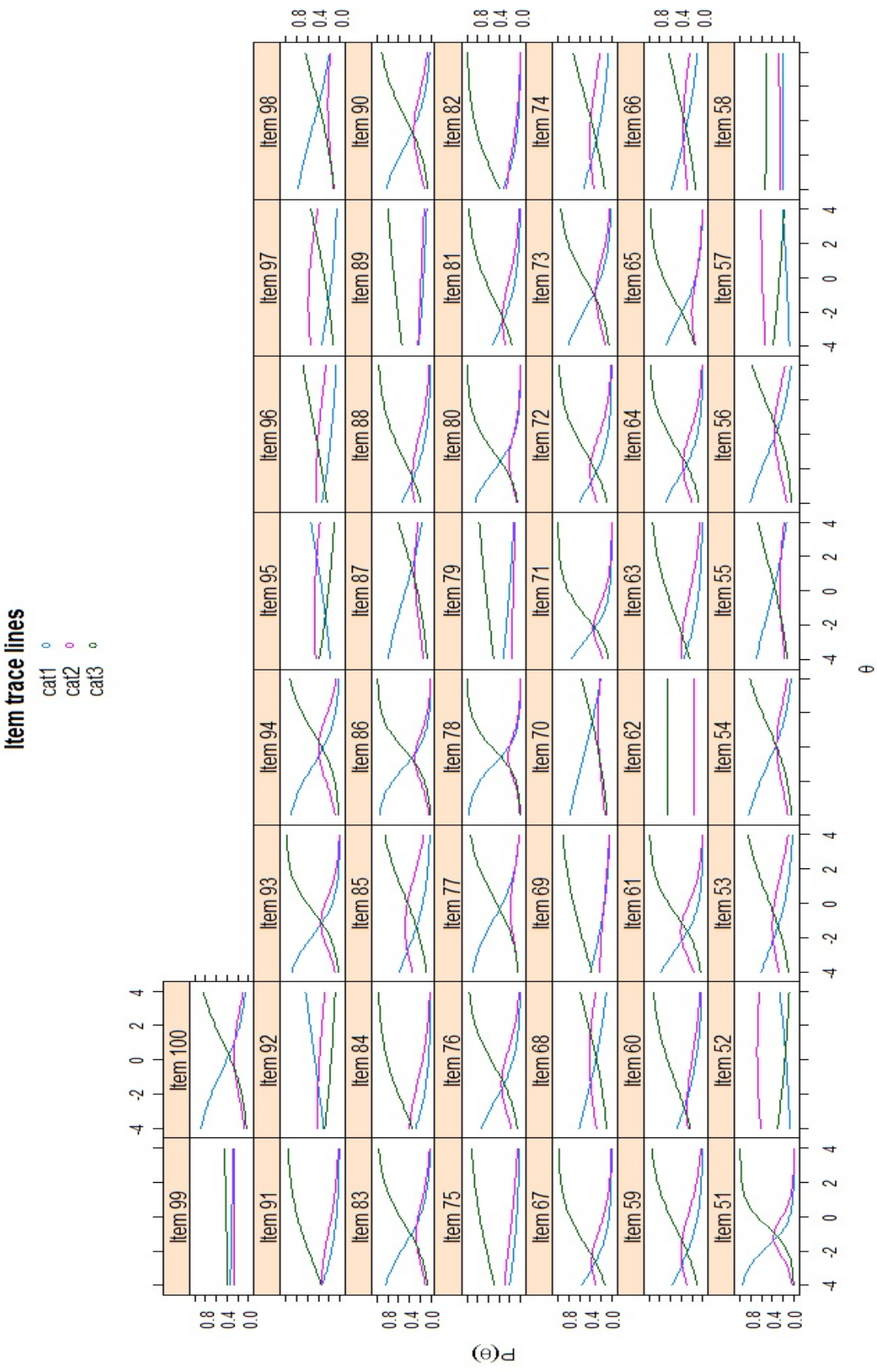

Figura 6.5: Curvas Características dos Itens 51 a 100 (Dados reais).

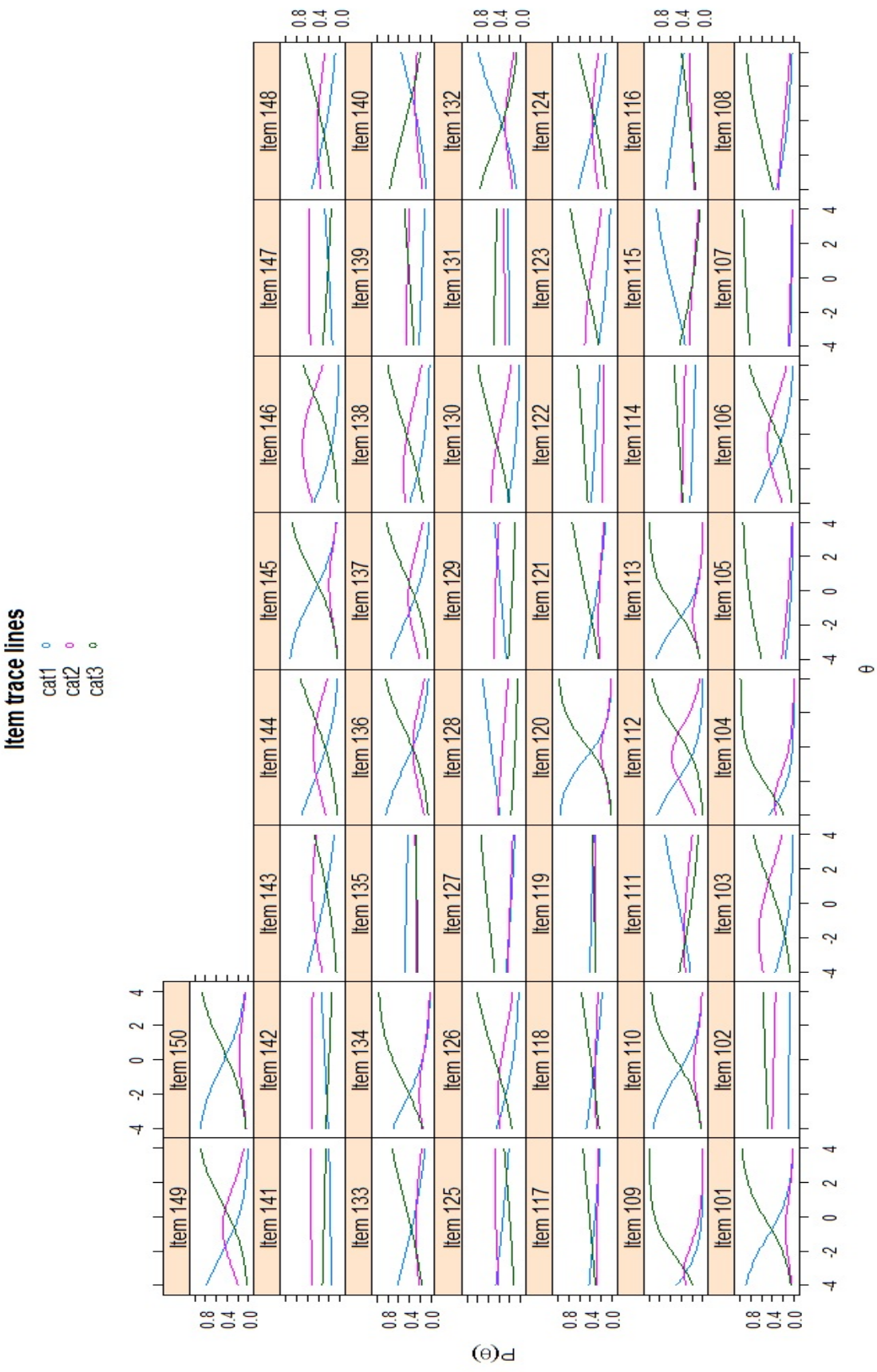

Figura 6.6: Curvas Características dos Itens 101 a 150 (Dados reais).

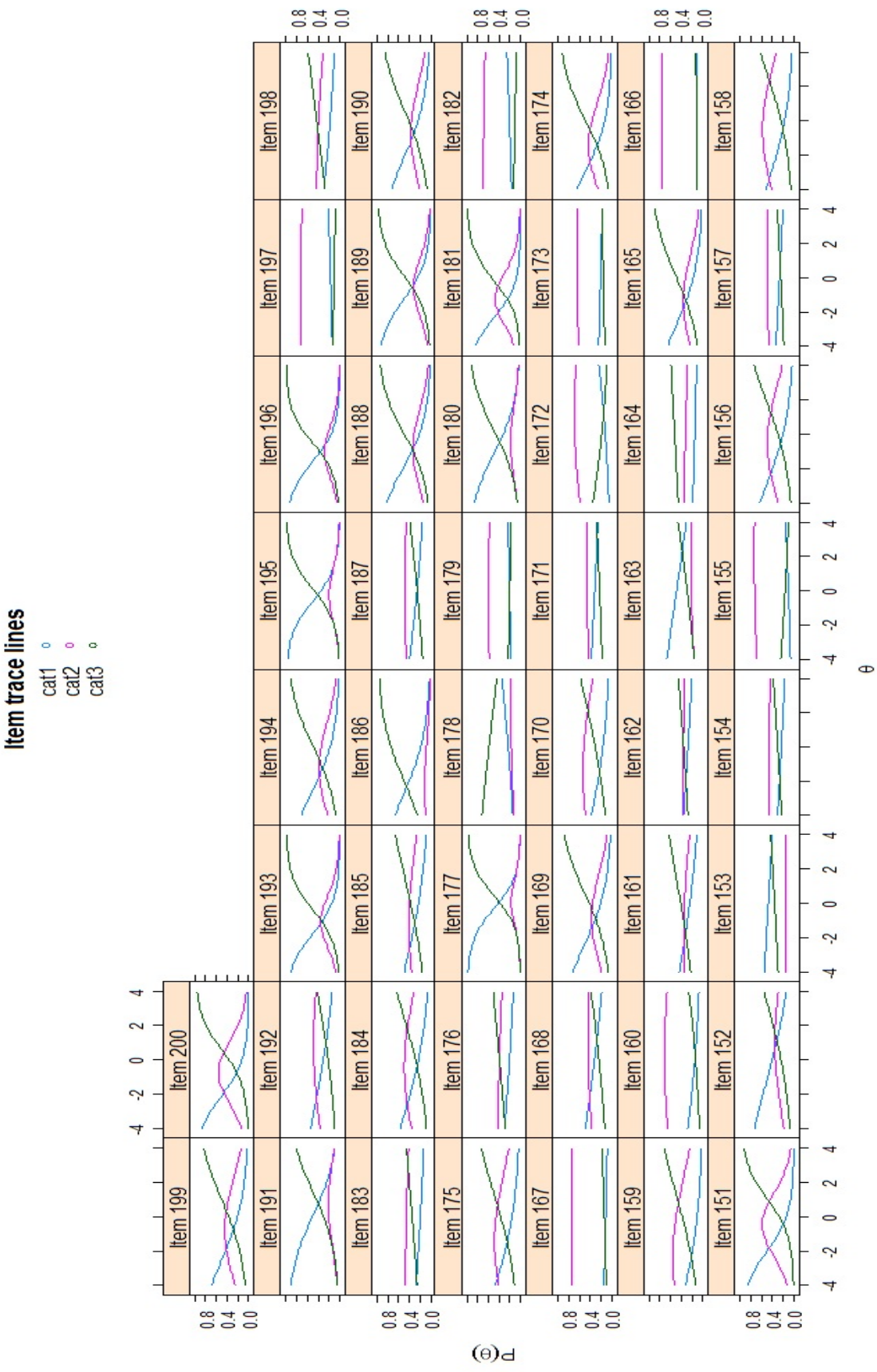

Figura 6.7: Curvas Características dos Itens 151 a 200. (Dados reais).

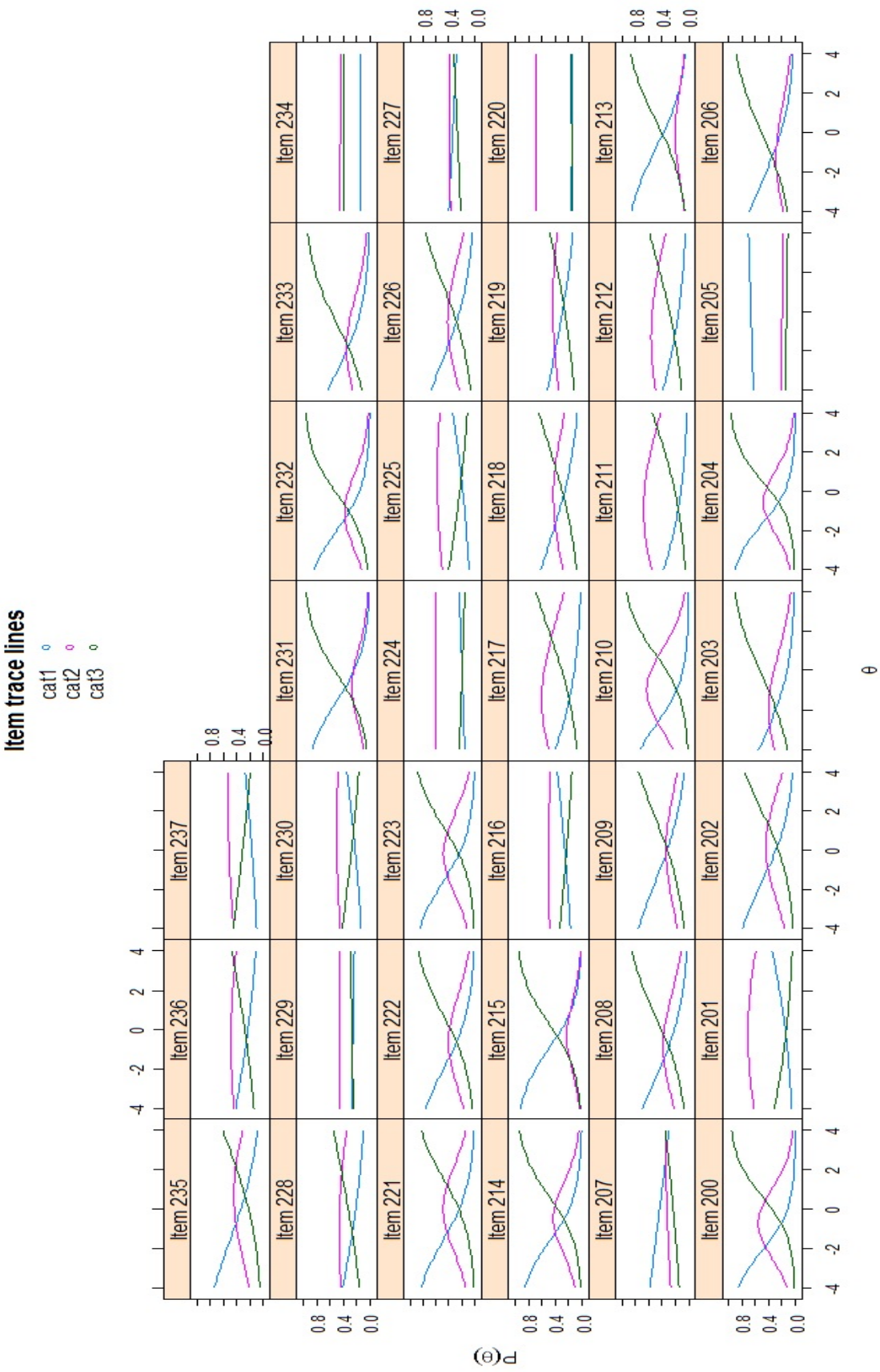

Figura 6.8: Curvas Características dos Itens 200 a 237 (Dados reais).

Na Figura 6.9, é apresentada a Curva de Informação do Teste (CIT). Através deste gráfico podemos observar que o instrumento de medida tem maior informação para os valores da habilidade compreendidos entre aproximadamente -4,8 a 2,5. Dessa forma, este resultado indica que a prova é mais propícia para avaliar habilidades com valores contidos nesse intervalo.

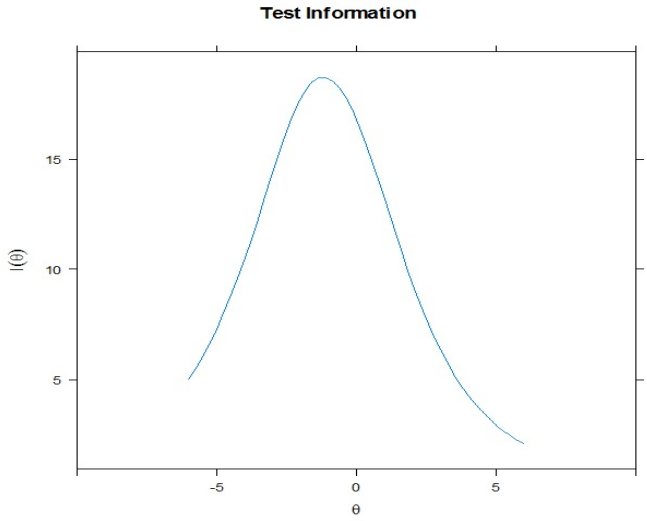

Figura 6.9: Curva Característica do Teste (Dados Reais).

6.3 Comparação entre as notas das provas corrigidas pelo Método Convencional e as notas das provas corrigidas pelo Modelo de Resposta Gradual

Conforme já apresentado anteriormente, o principal objetivo deste trabalho é comparar as notas dos indivíduos que prestam vestibular para a UnB, cujas provas são elaboradas pelo CEBRASPE, as quais são corrigidas pelo método convencional, com as notas obtidas através da correção pelo Modelo de Resposta Gradual (MRG). As questões analisadas neste estudo foram as questões do Tipo A, as quais são corrigidas através do método convencional, o qual se caracteriza pelo fato de o aluno ganhar a pontuação -1, 0 ou 1 caso ele erre, não responda ou acerte a questão, respectivamente.

De acordo com o mencionado na seção 5.3, há diferenças significativas entre os métodos de correção, visto que o método que segue o modelo de reposta gradual leva em consideração alguns pontos importantes que o método convencional não consegue captar.

Na Tabela 6.10, é apresentada uma análise descritiva das notas convencionais, das notas convencionais padronizadas e das notas estimadas através do MRG do banco de dados reais, que possui 237 questões e 7232 indivíduos. Da mesma forma como nos dados simulados, poderíamos supor que a menor nota convencional seria -237, caso houvesse algum aluno que errasse todas as questões. Entretanto, conforme vimos na seção 6.1, não houve nenhum respondente que errasse todas as questões, houve somente, um indivíduo que errou 78,9% da prova, o qual é o mesmo indivíduo que tirou a nota mínima, resultando em -139. Nota-se também que o máximo da nota convencional não resultou em 237, pois, como também mencionado na seção 6.1, não houve nenhum candidato que acertou todas as questões da prova. Dessa forma, o máximo da nota convencional, 178, foi atingido pelo indivíduo que acertou 83,12% da prova.

Tabela 6.10: *Análise Descritiva das Notas Convencionais, Notas Convencionais Padronizadas e Notas pelo MRG (Dados Reais).*

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Notas Convencionais	-139,000	19,000	37,000	41,470	60,000	178,000
Notas Convencionais Padronizadas	-5,741	-0,715	-0,142	0,000	0,590	4,344
Notas Modelo de Resposta Gradual	-2,981	-0,705	-0,198	0,000	0,511	3,872

Na Figura 6.10, as notas convencionais são apresentadas graficamente e é possível perceber que uma concentração das notas acima de zero, em torno de 40.

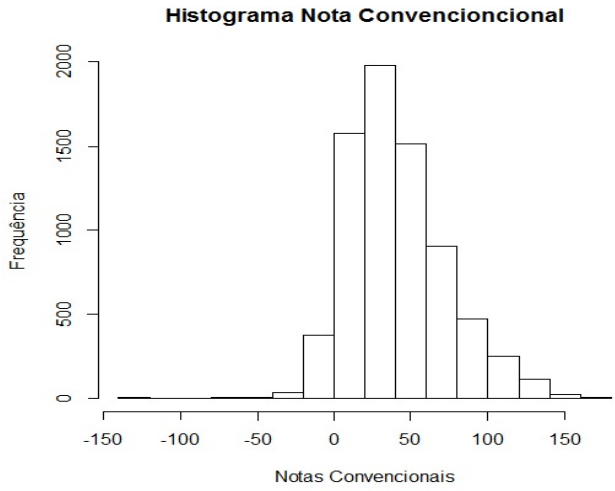

Figura 6.10: Histograma das Notas Convencionais (Dados Reais).

Ainda na Tabela 6.10, podemos observar a análise descritiva para as notas obtidas pelo modelo de resposta gradual. Comparando com os dados simulados, verificam-se valores aproximados em relação ao valor máximo das notas, e pouca diferença com relação à nota mínima. Como o MRG considera muito mais do que se o indivíduo respondeu de forma correta ou não a questão, não há valores padrão esperados das notas do MRG como, por exemplo, as quantidades de questões analisadas, conforme acontece nas notas corrigidas pelo método convencional.

Ainda na mesma tabela, podemos observar que os valores de máximo e mínimo das notas convencionais padronizadas não estão tão próximos das mesmas medidas das notas do MRG. Entretanto, o restante das medidas da análise descritiva está muito próximo.

Na Figura 6.11, são apresentados os histogramas das notas convencionais padronizadas e das notas do MRG. Podemos observar que as notas obtidas pelo MRG ficaram mais distribuídas em torno de -1 a 1 e que as notas convencionais padronizadas ficaram concentradas entre -1 e 0,5, com uma alta frequências nesse intervalo.

Com o intuito de realizar uma análise mais aprofundada, plotamos um gráfico de dispersão entre as notas convencionais padronizadas e as notas oriundas do modelo de resposta gradual, apresentadas no eixo X e no eixo Y, respectivamente, na Figura 6.12. Analisando este gráfico, podemos perceber uma tendência linear, como também foi visto nos dados simulados. Podemos confirmar também essa alta correlação mos-

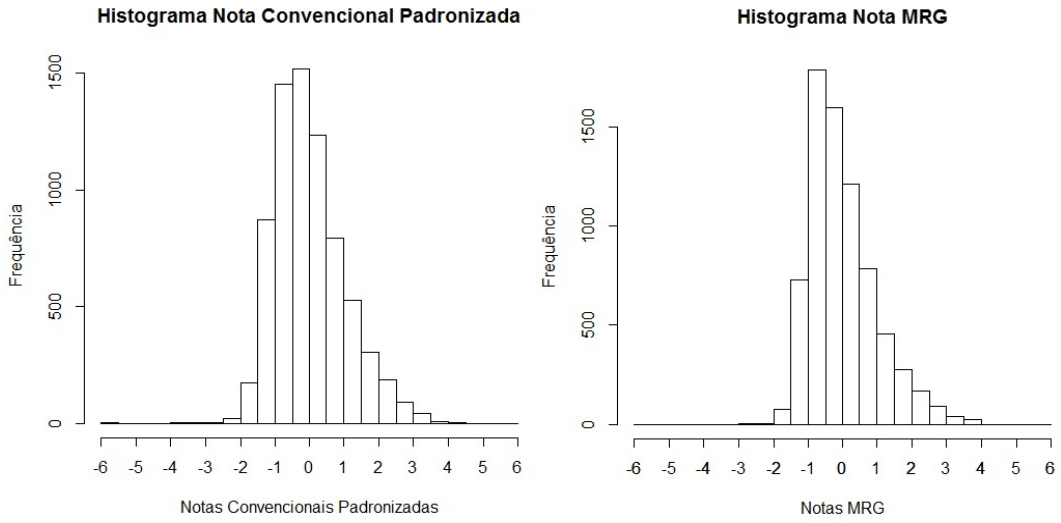

Figura 6.11: Histograma das Notas Convencionais Padronizadas e das Notas pelo MRG (Dados Reais).

trada pela linearidade dos dados no gráfico, através do valor da correlação entre as notas, o qual resultou em 0,9604.

Entretanto, comparando ainda com os dados simulados, podemos perceber uma tendência linear em praticamente toda a reta, ao contrário do que aconteceu com os dados simulados, os quais apresentaram uma alteração na linearidade na calda superior e inferior dos dados. Podemos supor, portanto, que a diferença nas notas mais elevadas e mais inferiores não tem tanta relevância ao se comparar as notas convencionais padronizadas e as notas vindas do MRG.

Analisando alguns indivíduos que obtiveram uma diferença significativa (como nos dados simulados, a diferença foi considerada significativa quando resultasse acima de 0,20) entre as notas obtidas pelos dois métodos, podemos observar nas Tabelas 6.11 e 6.12 as respostas de dois indivíduos às questões da prova e seus respectivos valores dos parâmetros dos itens. Ao analisarmos o indivíduo 6443, podemos notar que o fato de ele realmente não ter tido cautela ao responder praticamente todas as questões, errando assim muitas questões consideradas difíceis pelo modelo estimado, fez com que ele tivesse desvantagem na nota final do MRG, pois obteve um valor muito menor na nota pelo MRG em comparação com a nota convencional padronizada, sendo estas, -5,74 e -2,98, respectivamente.

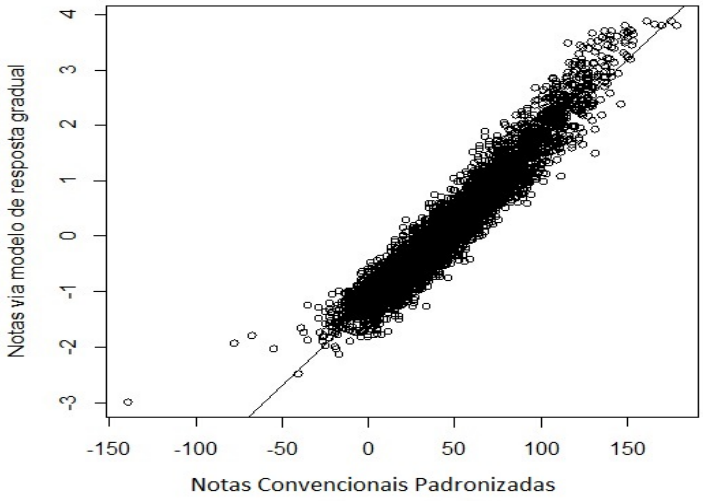

Figura 6.12: Diagrama de Dispersão entre as Notas Convencionais Padronizadas e as Notas pelo MRG (Dados Reais).

Da mesma forma, entretanto, com o efeito contrário, podemos observar que o indivíduo 4035, por ser muito cauteloso e parecer ter domínio do assunto medido, obteve um aumento considerável ao se comparar a nota convencional padronizada e a nota pelo MRG, sendo estas, 0,99 e 1,89, respectivamente. Podemos atribuir isso ao fato de que este indivíduo, ao não responder muitas questões por não ter certeza da resposta correta, deixou de errar talvez questões difíceis que iriam baixar sua nota no modelo de resposta Gradual. Do mesmo modo, por ter respondido poucas questões, mas de forma consciente, acertou tanto questões consideradas fáceis como difíceis pelo MRG, o que fez com que ele que ganhasse uma maior pontuação.

Tabela 6.11: *Respostas dos Indivíduos 6443 e 4035 (Itens 1 a 120-Dados Reais).*

	Parâmetros			Indivíduos			Parâmetros			Indivíduos			Parâmetros			Indivíduos	
Itens	a_i	b_{i1}	b_{i2}	6443	4035	Itens	a_i	b_{i1}	b_{i2}	6443	4035	Itens	a_i	b_{i1}	b_{i2}	6443	4035
1	0,407	0,595	2,38	-1	0	41	0,68	-2,895	-1,214	1	1	81	0,613	-3,648	-1,337	-1	1
2	-0,097	-4,643	-13,436	-1	1	42	0,363	-4,629	-2,01	1	1	82	0,583	-5,117	-3,302	-1	1
3	0,156	-6,417	-1,251	1	0	43	0,809	-2,954	-0,474	-1	0	83	0,764	-1,794	-0,33	1	0
4	0,205	-0,974	5,484	-1	0	44	0,15	-5,413	7,012	-1	0	84	0,541	-5,734	-2,651	0	1
5	0,806	-1,709	-0,584	-1	1	45	0,176	-9,625	-1,233	1	0	85	0,524	-3,337	0,622	-1	0
6	-0,127	-1,458	-7,126	-1	0	46	0,534	-3,995	0,612	-1	0	86	1,089	-1,217	-0,098	-1	1
7	0,1	-1,088	13,523	-1	0	47	0,09	-24,547	-4,976	-1	1	87	0,375	-0,396	2,901	-1	-1
8	-0,235	-1,426	-4,971	-1	0	48	0,356	-3,99	1,911	1	0	88	0,603	-3,849	-1,447	1	1
9	0,515	-4,383	-2,834	-1	1	49	0,847	-0,404	0,207	-1	1	89	0,153	-12,196	-5,034	1	1
10	0,674	-2,825	-0,724	1	0	50	0,075	-10,769	13,554	-1	1	90	0,649	-1,609	0,41	-1	0
11	0,701	-3,745	-2,122	1	1	51	1,374	-1,743	-0,568	-1	1	91	0,496	-5,478	-2,61	-1	1
12	0,293	-1,301	1,623	-1	0	52	-0,192	9,199	-8,22	-1	0	92	-0,175	0,862	-9,5	-1	0
13	0,477	-3,087	-0,138	-1	0	53	0,497	-3,02	0,403	-1	1	93	0,998	-1,882	-0,412	1	1
14	0,599	-1,987	-0,076	1	0	54	0,567	-1,053	1,315	-1	0	94	0,793	-1,153	0,927	1	0
15	0,646	-1,849	-0,162	-1	0	55	0,335	-1,393	1,81	-1	0	95	-0,211	3,374	-6,37	1	0
16	0,142	4,498	9,564	-1	0	56	0,538	-1,242	1,647	-1	1	96	0,239	-7,054	0,856	-1	0
17	-0,138	2,411	-7,798	-1	0	57	-0,13	14,192	-7,65	-1	0	97	0,268	-6,483	3,463	-1	0
18	0,464	-2,384	0,498	-1	0	58	-0,006	222,627	16,219	1	1	98	0,341	-0,343	2,287	-1	1
19	0,334	2,276	4,692	-1	0	59	0,633	-3,445	-0,721	1	1	99	0,032	-24,463	10,198	-1	0
20	0,156	-10,963	3,247	-1	0	60	0,46	-4,128	-1,463	-1	1	100	0,598	-0,671	1,219	1	0
21	0,718	-2,014	-0,402	1	0	61	0,899	-2,612	-0,663	-1	1	101	0,768	-1,161	-0,342	1	1
22	0,687	-0,722	0,927	-1	0	62	-0,004	375,234	163,139	1	1	102	0,047	-49,516	-2,971	1	0
23	1,093	-2,228	-1,268	1	0	63	0,472	-5,269	-1,65	-1	1	103	0,461	-5,287	1,492	-1	0
24	0,03	-45,263	-0,37	-1	0	64	0,781	-2,807	-0,807	-1	1	104	0,841	-4,15	-2,314	1	1
25	0,455	-0,477	0,847	1	0	65	0,696	-2,682	-1,488	1	1	105	0,308	-9,305	-5,422	1	1
26	0,381	-5,54	-0,12	1	0	66	0,309	-2,953	2,016	-1	1	106	0,597	-2,316	1,271	-1	0
27	0,602	-3,988	-0,633	1	0	67	0,793	-3,657	-1,651	-1	1	107	0,172	-18,817	-12,868	1	1
28	0,539	-4,803	-2,461	1	0	68	0,327	-2,592	2,814	-1	1	108	0,318	-6,337	-2,581	0	1
29	1,023	-1,742	-0,618	-1	0	69	0,36	-5,077	-2,624	1	1	109	0,925	-4,021	-2,426	1	1
30	0,346	-1,647	0,235	1	-1	70	0,325	-0,129	3,107	-1	1	110	0,804	-0,895	-0,04	1	1
31	0,871	-1,744	-0,717	1	0	71	1,036	-2,889	-1,485	1	1	111	-0,263	0,598	-4,882	-1	0
32	-0,639	-2,334	-3,381	-1	-1	72	0,753	-3,331	-1,036	1	1	112	0,886	-2,023	0,945	-1	1
33	1,595	-1,967	-1,354	1	1	73	0,713	-1,825	-0,087	-1	1	113	0,939	-1,922	-1,117	-1	1
34	0,715	-1,918	-0,151	1	1	74	0,353	-3,618	1,31	1	1	114	0,082	-18,32	2,248	-1	1
35	0,214	-2,508	1,075	-1	1	75	0,307	-8,309	-3,859	1	1	115	-0,318	-1,656	-4,695	-1	1
36	0,452	-3,764	-1,885	1	0	76	0,718	-2,425	-0,344	-1	1	116	0,176	0,652	6,216	-1	-1
37	0,966	-1,179	-0,707	-1	1	77	0,706	-0,83	0,276	-1	1	117	0,122	-6,45	2,713	-1	0
38	0,701	-2,824	0,162	-1	0	78	1,214	-0,978	-0,167	-1	1	118	0,179	-4,564	2,456	-1	0
39	0,447	-0,961	1,751	-1	0	79	0,151	-8,757	-4,126	1	1	119	0,037	-14,059	19,841	-1	0
40	0,885	-3,831	-2,255	1	0	80	0,949	-2,065	-1,11	1	1	120	1,015	-0,651	0,182	-1	1

Tabela 6.12: *Respostas dos Indivíduos 6443 e 4035 (Itens 120 a 237-Dados Reais).*

	Parâmetros			Indivíduos			Parâmetros			Indivíduos			Parâmetros			Indivíduos	
Itens	a_i	b_{i1}	b_{i2}	6443	4035	Itens	a_i	b_{i1}	b_{i2}	6443	4035	Itens	a_i	b_{i1}	b_{i2}	6443	4035
121	0,268	-3,62	0,098	-1	0	161	0,223	-5,034	1,575	-1	0	201	-0,275	6,193	-6,757	-1	0
122	0,095	-8,968	-1,696	-1	0	162	0,096	-9,894	5,952	-1	0	202	0,559	-1,548	1,908	-1	0
123	0,296	-7,569	-0,166	-1	0	163	0,187	-0,224	4,721	-1	1	203	0,531	-3,575	-0,331	-1	1
124	0,333	-2,275	2,28	-1	0	164	0,069	-25,215	-1,521	-1	1	204	0,982	-1,651	0,5	-1	1
125	0,14	-5,056	9,634	-1	0	165	0,545	-2,792	0,017	-1	0	205	-0,041	-16,921	-45,281	-1	1
126	0,385	-4,497	0,166	-1	0	166	0,02	-105,846	99,746	-1	0	206	0,511	-2,437	-0,101	-1	0
127	0,142	-10,755	-3,595	-1	0	167	0,082	-25,505	22,385	-1	0	207	0,148	-1,675	8,028	-1	1
128	-0,168	-1,336	-12,832	-1	0	168	0,179	-4,179	6,484	-1	0	208	0,533	-2,397	0,665	-1	1
129	-0,118	4,295	-14,222	-1	0	169	0,574	-2,279	0,499	-1	1	209	0,463	-1,53	1,569	-1	1
130	0,342	-7,141	-0,079	-1	0	170	0,286	-5,629	2,796	-1	0	210	0,837	-2,756	0,804	-1	1
131	-0,033	37,455	-2,926	-1	0	171	0,073	-10,521	16,812	-1	0	211	0,378	-5,269	3,468	-1	0
132	-0,468	1,061	-1,461	-1	0	172	-0,215	9,242	-6,793	-1	0	212	0,313	-5,412	2,784	-1	1
133	0,323	-2,519	1,068	-1	0	173	0,051	-26,145	33,287	-1	0	213	0,585	-0,816	0,607	-1	1
134	0,63	-2,661	-1,289	-1	0	174	0,637	-2,901	0,07	-1	1	214	0,819	-1,65	0,622	-1	1
135	0,03	-5,53	35,107	-1	0	175	0,385	-4,26	1,296	-1	1	215	0,812	-0,922	0,257	-1	1
136	0,582	-1,224	1,166	-1	0	176	0,12	-11,178	3,549	-1	1	216	-0,128	8,119	-9,304	-1	1
137	0,555	-2,037	1,205	-1	0	177	1,069	-0,254	0,422	-1	1	217	0,41	-4,903	1,961	-1	1
138	0,396	-5,191	0,494	-1	0	178	-0,147	8,076	2,535	-1	-1	218	0,376	-2,674	2,241	-1	1
139	0,095	-17,137	4,284	-1	0	179	-0,042	32,25	-32,875	-1	0	219	0,243	-3,524	4,235	-1	0
140	-0,324	3,343	-0,443	-1	0	180	0,637	-1,092	0,129	-1	-1	220	0,029	-59,052	57,684	-1	0
141	-0,038	40,323	-23,142	-1	1	181	1,019	-2,32	-0,259	-1	1	221	0,658	-1,618	1,601	-1	0
142	-0,083	11,677	-16,838	-1	1	182	-0,065	20,589	-33,141	-1	0	222	0,612	-2,079	0,773	-1	1
143	0,32	-2,669	4,364	-1	1	183	0,111	-12,791	5,707	-1	0	223	0,736	-1,617	1,238	-1	0
144	0,5	-2,263	2,023	-1	1	184	0,371	-3,279	2,576	-1	1	224	-0,066	21,053	-21,223	-1	0
145	0,661	-0,271	0,938	-1	1	185	0,295	-4,098	1,797	-1	0	225	-0,198	7,284	-5,937	-1	0
146	0,506	-4,198	2,531	-1	1	186	0,538	-2,761	-1,892	-1	1	226	0,462	-2,421	1,438	-1	0
147	-0,11	12,713	-11,114	-1	0	187	0,158	-6,755	6,84	-1	1	227	0,064	-10,148	15,357	-1	0
148	0,311	-3,7	2,038	-1	1	188	0,692	-1,952	0,087	-1	1	228	0,232	-5,611	3,146	-1	0
149	0,686	-2,196	0,849	-1	1	189	0,917	-1,33	0,123	-1	0	229	0,025	-42,404	39,11	-1	0
150	0,605	-0,327	0,822	-1	-1	190	0,537	-2,219	0,713	-1	1	230	-0,155	8,021	-6,168	-1	0
151	0,91	-1,961	1,039	-1	1	191	0,58	0,01	1,493	-1	1	231	0,76	-1,581	-0,119	-1	1
152	0,345	-1,054	3,293	-1	0	192	0,245	-3,355	5,352	-1	0	232	0,826	-2,021	-0,08	1	0
153	0,075	-1,175	7,651	-1	0	193	1,026	-1,781	-0,342	-1	1	233	0,58	-3,104	-0,569	-1	0
154	0,092	-12,889	9,095	-1	0	194	0,609	-2,553	0,116	-1	1	234	0,001	-2999,52	637,359	1	0
155	-0,132	16,487	-12,427	-1	0	195	1,036	-0,655	0,121	-1	1	235	0,416	-1,583	3,01	-1	1
156	0,474	-2,849	1,76	-1	1	196	1,054	-1,629	-0,508	-1	1	236	0,211	-5,616	4,487	1	1
157	0,084	-12,016	13,36	-1	0	197	-0,063	24,949	-33,883	-1	0	237	-0,15	10,517	-5,31	-1	0
158	0,418	-3,733	2,677	-1	0	198	0,167	-9,171	1,83	-1	0						
159	0,35	-6,023	1,275	-1	0	199	0,534	-2,552	1,026	-1	1						
160	0,203	-8,594	9,114	-1	0	200	0,936	-2,094	0,622	-1	1						

Conforme apresentado no capítulo anterior, foi realizado também uma análise dos ranks a partir da análise de Regressão de Nadaraya-Watson, com o intuito de "deixar os dados mostrarem por si só qual a curva que melhor se ajusta", a qual representará a relação entre eles. Ou seja, através dessa técnica, assume-se que a relação funcional entre as variáveis é completamente desconhecida e usam-se os dados para inferir sobre a sua forma.

Analisando a Figura 6.13, similar à Figura 5.10 do capítulo anterior, podemos notar indícios, agora a partir de dados reais, também de uma aproximação entre os ranks obtidos pelo método convencional e pelo método MRG para indivíduos que obtiveram notas maiores, ocupando as primeiras classificações. E ainda, diferente do que ocorreu nos dados simulados, há indícios também de um alto grau de semelhança entre os ranks dos indivíduos que obtiveram uma colocação mais baixa, correspondendo aos respondentes que obtiveram notas menores. Por fim, podemos observar também que há indícios de diferenças entre os ranks para os indivíduos que ficaram com classificação entre os primeiros colocados e os últimos, diferente também do que ocorreu com os dados simulados.

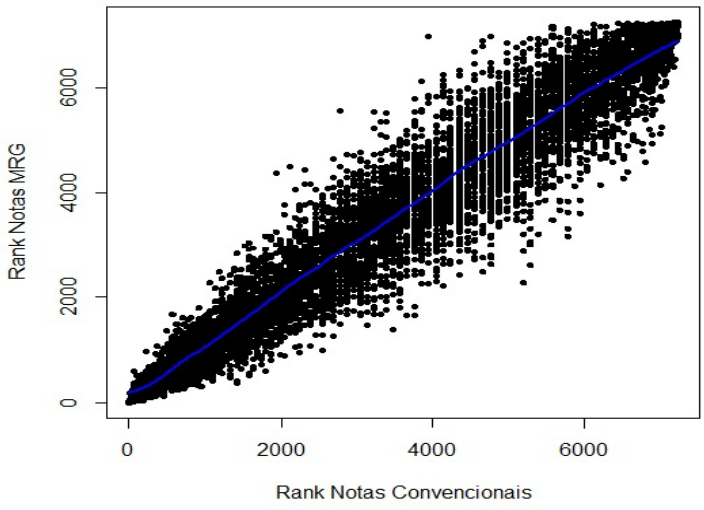

Figura 6.13: Curva estimada pelo método de Nadaraya-Watson entre os Ranks obtidos pelo método MRG e pelo método Convencional (Dados Reais).

Na Figura 6.14, há um gráfico em que estão plotados o módulo das diferenças entre os ranks das notas obtidas pelo método convencional e pelo método MRG no

eixo Y e o rank da notas obtidas pelo método convencional no eixo X. A partir deste gráfico, é possível perceber de forma mais evidente a grande diferença entre os ranks obtidos pelo método convencional e MRG situados na parte central da classificação dos indivíduos e uma conformidade entre os ranks dos indivíduos com alta e baixa classificação no teste.

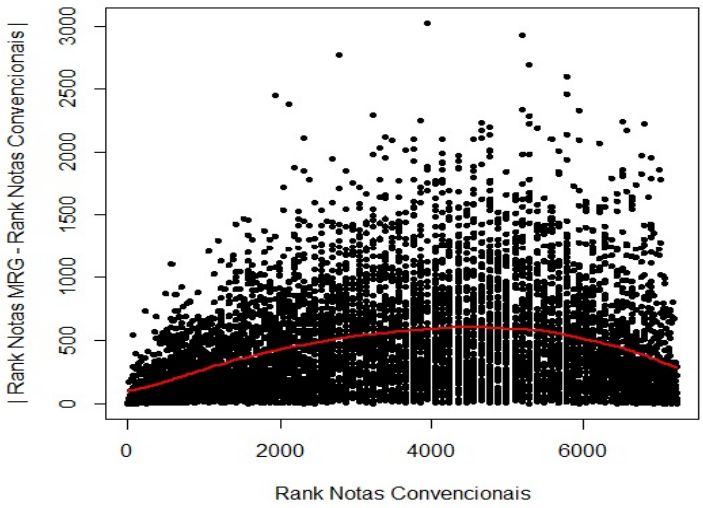

Figura 6.14: Curva estimada pelo método de Nadaraya-Watson entre o módulo da diferença entre os Ranks obtidos pelo método MRG e pelo método Convencional e Rank Convencional (Dados Simulados).

Na Figura 6.15 está apresentado o gráfico em que mostra a relação entre o grau de discordância dos ranks obtidos pelo método de correção convencional e pelo método de correção proposto, modelo de resposta gradual, em função do número de candidatos selecionados, utilizando os 7232 indivíduos e os 237 itens analisados. Podemos perceber um alto grau de discordância para um número baixo de número de indivíduos selecionados e, a partir de aproximadamente 1000 indivíduos, esse grau de discordância começa a se estabilizar até em torno de 3000 indivíduos e, em seguida, há um alto grau de discordância para os indivíduos que ficaram nas últimas colocações.

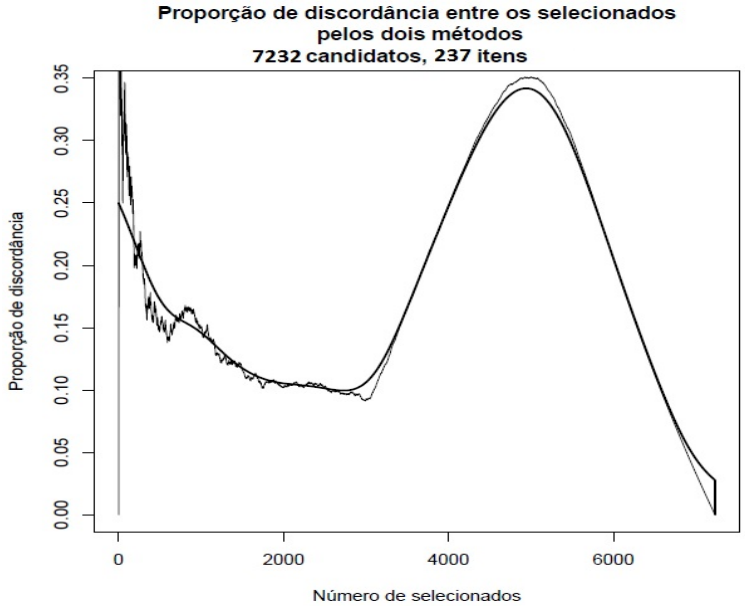

Figura 6.15: Gráfico de discordância entre o método convencional e MRG (Dados Reais).

Em específico, se por acaso fossem selecionados 1000 indivíduos, dentre esses 7232 candidatos analisados, aproximadamente 15% dos indivíduos que ficaram entre os 1000 primeiros candidatos selecionados pelo método convencional, não estariam entre os 1000 primeiros candidatos selecionados pelo MRG. Em outras palavras, em torno de 150 candidatos, dentre os 1000 selecionados pelo método convencional, não seriam selecionados pelo método MRG. E os outros 150 candidatos selecionados pelo método MRG estariam além dos 1000 primeiros candidatos selecionados pelo método convencional, classificados fora das vagas. Podemos atribuir esse fato a diferença de avaliação entre os métodos, em que o MRG considera se os indivíduos acertaram questões fáceis ou difíceis. Baseado nisso, podemos supor que estes indivíduos que foram selecionados pelo método MRG tenham acertado questões mais difíceis na prova, fazendo com que ficassem dentro da classificação do MRG e fora da classificação do método convencional. Podemos considerar um grau de discordância entre os métodos de 15% baixo, concluindo, assim, que o modelo de resposta gradual se aproxima com qualidade em relação a estimação das notas, com vantagens adicionais.

Dessa forma, do mesmo modo que observamos na situação em que os dados foram simulados, podemos concluir que o modelo de resposta gradual da TRI também estima notas dos indivíduos submetidos ao vestibular da UnB de forma aproximada

ao método convencional de correção utilizado pelo CEBRASPE, para os indivíduos com maior classificação. Assim, podemos concluir que o modelo de resposta gradual funciona de forma eficiente para diferentes situações, com diferente número de respondente de uma prova, diferente número de questões, tanto numa situação simulada como numa situação real.

Desta maneira, podemos ressaltar as vantagens em se realizar a correção das provas do vestibular da UnB através do MRG, por pertencer à família de modelos da técnica de Teoria de Resposta ao Item, a qual avalia de forma mais adequada os indivíduos, considerando a dificuldade do item do teste e, também, discriminando de forma diferente indivíduos que tenham um maior grau de conhecimento no assunto medido, fazendo com que indivíduos como este tenham benefícios por pelo alto conhecimento. Outra vantagem que podemos citar também do método proposto de correção é a comparação entre indivíduos submetidos a provas completamente diferentes com algumas questões similares, fazendo com que, por exemplo, seja possível a avaliação de melhora de rendimento ou conhecimento entre um ano em questão e seu posterior, de alunos que tenham determinado perfil e conhecimento sobre algum traço latente.

Capítulo 7

Conclusões e Trabalhos Futuros

7.1 Conclusões

Neste trabalho, foi realizada uma análise das notas de alunos submetidos à uma prova do vestibular da UnB, elaborada pelo CEBRASPE. Essa análise constituiu-se, principalmente, de uma comparação entre as notas destes participantes e, consequentemente, uma comparação também da ordem de classificação entre os mesmos, obtidas pelo método de avaliação proposto, correção pelo modelo de resposta gradual da TRI, e obtidas pelo método convencionalmente realizado pelo CEBRASPE. Para isso, foram analisadas duas situações: Uma simulada, gerada através de regras definidas, e uma situação real, gerada a partir de um banco de dados real disponibilizado pelo CEBRASPE.

Ao analisarmos os perfis dos respondentes das provas e compararmos o comportamento dos dados simulados e dos dados reais, foi possível perceber que obtivemos uma situação bastante próxima da realidade nos dados simulados, em que conseguimos adicionar em nossa base simulada diferentes perfis de respondentes, com diferentes comportamentos, se aproximando assim da realidade. Entretanto, quando se analisou a distribuição de respostas global entre as categorias, obtivemos nos dados simulados um número maior de não respostas do que nos dados reais. Para este último, foi possível perceber, ainda, um parcial equilíbrio de distribuição de escolhas entre as categorias de respostas.

Com relação à análise dos parâmetros dos itens estimados pelo Modelo de

Resposta Gradual (MRG), nos dados simulados, observamos que, em geral, os parâmetros dos itens estavam dentro do intervalo adequado e de acordo com a distribuição de respostas de cada categoria e com suas respectivas CCIs. Entretanto, foi obervado também, valores dos parâmetros de alguns itens não esperados, fora do intervalo ideal, gerando assim estimativas ruins. Para os dados reais, com um número bem maior de indivíduos e questões avaliadas, foram observados também itens com parâmetros bem estimados, assim como também suas respectivas CCIs, mas, também, entretanto, alguns itens com parâmetros mal estimados, representando itens ruins no teste.

Com relação a comparação entre as notas, nos dados simulados, foi observada uma proximidade entre as notas obtidas pelo método convencional e pelo MRG e também uma alta correlação entre as mesmas, de forma positiva. Através da utilização da Regressão de Nadaraya-Watson, essa relação de linearidade entre as notas foi evidenciada, indicando assim uma real proximidade e concordância, de forma geral, entre as notas obtidas pelo método convencional e pelo modelo de resposta gradual. Para alguns casos de não concordância, foi observado que estes só ocorreram em notas mais baixas, ou seja, na classificação de indivíduos fora da região de interesse, pois a quantidade de vagas é limitada no vestibular. Ainda no sentido de comparação das notas, os dados reais se comportaram de forma semelhante, resultando em uma alta correlação entre as notas convencionais padronizadas e as notas obtidas pelo MRG, com algumas pequenas diferenças irrelevantes dos dados simulados. A relação de linearidade também foi evidenciada através da Regressão de Nadaraya-Watson, juntamente com o alto valor de correlação entre as notas.

Dessa forma, podemos concluir que os resultados finais demonstraram a potencialidade do método de correção proposto em estimar as notas dos indivíduos que são submetidos às questões do tipo A das provas do vestibular da UnB, mostrando resultados eficientes. Além disso, foi possível perceber a competência do modelo em estimar as notas em situações diferentes com, por exemplo, número de indivíduos e itens distintos. Assim, podemos constatar que o método proposto pode representar uma nova forma de correção das provas por conter vantagens sobre o método comumente utilizado, tais como: comparar indivíduos de populações diferentes, quando submetidos a testes ou provas que tenham somente alguns itens em comum; analisar indivíduos de forma mais adequada, considerando, por exemplo, a dificuldade dos

itens e também o quanto ele pode discriminar indivíduos que possuam ou não um alto conhecimento do traço latente medido.

7.2 Trabalhos Futuros

Apesar de o modelo de resposta gradual da Teoria da resposta ao item ter se mostrado eficiente em estimar as notas dos indivíduos analisados, diversas sugestões surgem a partir das análises feitas, gerando várias outras oportunidades de trabalhos dentro do cenário abordado. Dentre elas, podemos citar:

- Fazer análises de outros anos do vestibular da UnB, e comparar com os resultados obtidos neste trabalho, observando se há ou não um comportamento semelhante.

- Elaborar uma forma de realizar a proposta deste trabalho levando em consideração as disciplinas/áreas abordadas no teste e também levar em consideração o curso para o qual o indivíduo se inscreveu.

- Realizar o mesmo procedimento excluindo os itens que se mostraram não informativos através dos valores estimados dos parâmetros dos itens, para analisar mudanças ou não nas notas estimadas.

- Propor um método de avaliação também para as questões de tipo C, a qual constitui um item de múltipla escolha.

- Elaborar um método alternativo de imputação dos dados via Regressão Logística para as questões deixadas em branco, com o objetivo de se estimar qual a probabilidade de o indivíduo responder corretamente a questão, baseado no perfil de acertos e erros das outras questões respondidas pelos indivíduos, obtendo assim, um banco de dados completos.

Referências Bibliográficas

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, 67(4):1012–1028.

Alexandre, J. W. C., Andrade, D. d., Vasconcelos, A. d., Araujo, A. d., & Batista, M. J. (2003). Análise do número de categorias da escala de likert aplicada à gestão pela qualidade total através da teoria da resposta ao item. *Encontro Nacional De Engenharia De Produção*, 23:1–20.

Alexandre, J. W. C., Andrade, D. F. d., Vasconcelos, A. P. d., & Araújo, A. M. S. d. (2002). Uma proposta de análise de um construto para medição dos fatores críticos da gestão pela qualidade por intermédio da teoria da resposta ao item. *Gestão e Produção*, 9(2):129–141.

Andrade, D. F. d., Tavares, H. R., & da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*.

Andrade de, D. F. (2005). *Teoria da resposta ao item: Conceitos, Modelos e Aplicações*. PhD thesis, IME, Universidade de São Paulo.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.

Andriola, W. B. (2009). Psicometria moderna: características e tendências. *Estudos em Avaliação Educacional*, 20(43):319–340.

Araujo, E. A. C. d., Andrade, D. F. d., & Bortolotti, S. L. V. (2009). Teoria da resposta ao item. *Revista da Escola de Enfermagem USP, São Paulo*, 43:1000–1008.

Assunção, F. (2012). *Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos*. PhD thesis, Universidade de São Paulo.

Azevedo, C. L. N. (2003). *Métodos de estimação na teoria de resposta ao item*. PhD thesis, Instituto de Matemática e Estatística da Universidade de São Paulo, 27/02/2003.

Baker, F. B. (1992). *Item Response Theory- Parameter Estimation Techniques*. New YorK: Marcel Dekker, Inc.

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51.

Bock, R. D. & Zimowski, M. F. (1997). Multiple group irt. In: *Handbook of modern item response theory*, pages 433–448. Springer.

Bosi, M. (2010). Um estudo sobre o grau de maturidade e a evolução da gestão pela qualidade total no setor de transformação cearense por meio da teoria da resposta ao item. 2010. 135f. *Universidade Federal do Ceará, Curso de Mestrado em Logística e Pesquisa Operacional, Fortaleza*.

Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189.

Cai, Z. (2001). Weighted nadaraya–watson regression estimation. *Statistics & probability letters*, 51(3):307–318.

Cortes, R. X. (2010). Um estudo comparativo de estimadores de regressoes não-paramétricas aditivas: Performance em amostras finitas. Instituto de Matemática, Departamento de Estatística, Universidade Federal do Rio Grande do Sul (UFRG).

Demars, C. (2010). *Item response theory*. Oxford University Press, New York.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Ferreira, E. V. (2014). Modelos da teoria de resposta ao item assimétricos de grupos múltiplos para respostas politômicas nominais e ordinais sob um enfoque bayesiano. *Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica*.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576.

Guewehr, K. (2007). *Teoria da resposta ao item na avaliação de qualidade de vida de idosos*. PhD thesis, Dissertação de mestrado-Universidade Federal do Rio Grande do Sul, Faculdade de Medicina, Programa de Pós-Graduação em Epidemiologia, RS.

Guilford, J. P. (1954). Psychometric methods. McGraw Hill, New York.

Gulliksen, H. (1950). Theory of mental tests. John Wiley and Sons, New York.

Hunter, D. R., Richards, D. S. P., & Rosenberger, J. L. (2011). *Nonparametric Statistics and Mixture Models*. World Scientific.

Junior, F. d. J. M., Zanella, A., Lopes, L. F. D., & Seidel, E. J. (2015). Avaliação da satisfação de alunos por meio do modelo de resposta gradual da teoria da resposta ao item. *Revista Ensaio: Avaliação e Políticas Públicas em Educação*, 23(86):129–158.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. In: *American Political Science Association*, volume 95, pages 49–69. Cambridge Univ Press.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Lord, F. (1952). A theory of test scores. *Psychometric monograph*. Vol. 7.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates, Inc.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, 1992(1):i–30.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Nunes, L. N. (2007). *Métodos de imputação de dados aplicados na área da saúde*. PhD thesis. Faculdade de Medicina, Programa de Pós-Graduação em Epidemologia, Universidade Federal do Rio Grande do Sul (UFRG).

Pereira, E. A. (2014). Algumas propostas para imputação de dados faltantes em teoria de resposta ao item. Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília (UnB).

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Santos, V. L. F. & dos, G. (2009). Teoria de resposta ao item: uma abordagem generalizada das curvas caracterısticas dos itens. *Universidade Federal do Rio de Janeiro*.

Schafer, J. L. & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Silva, A. d. J. M. d. (2010). Estimação da média com observações em falta. Master's thesis, Universidade de Coimbra.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Siminassi, A. G. & Júnior, J. O. C. (2005). Econometria não paramétrica e expectativa de vida nos municípios do nordeste: uma aplicação do estimador de nadaraya-watson. *Fórum bnb de desenvolvimento - encontro regional de economia da ANPEC*, 10.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Soares, T. M. (2005). Utilização da teoria da resposta ao item na produção de indicadores sócio-econômicos. *Pesquisa Operacional*, 25(1):83–112.

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer Science & Business Media.

Veroneze, R. (2011). *Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla*. PhD thesis, Universidade Estadual de Canpinas.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In: *Proceedings of the 1967 invitational conference on testing problems*, pages 85–101. Educational Testing Service Princeton, NJ.