Universidade de Brasília
Departamento de Estatística

Avaliação das Notas das Questões de Química do Vestibular da UnB via Teoria de Resposta ao Item

Otávio Augusto Morosini Mansur de Carvalho

Orientador: Antonio Eduardo Gomes

Brasília
Maio de 2022

Otávio Augusto Morosini Mansur de Carvalho

Avaliação das Notas das Questões de Química do Vestibular da UnB via Teoria de Resposta ao Item

Orientador: Antonio Eduardo Gomes

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
Maio de 2022

"São várias as maneiras de você tomar uma decisão, e uma dessas maneiras é você tomar a decisão baseada em dados. Não é a melhor, não é a pior é apenas uma delas."

-Wilton de Oliveira Bussab

Resumo

O trabalho consiste em avaliar as 38 questões de química do vestibular da UnB de 2014, utilizando o escore convencional obtido pelos candidatos e modelos de Teoria da Resposta ao Item para identificar questões com baixo poder de discriminação entre candidatos com diferentes níveis de conhecimento em química. Serão apresentados modelos de Teoria da Resposta ao Item, como os Modelos Dicotômicos de dois e três parâmetros (ML2 e ML3) e o Modelo de Resposta Gradual (MRG), como alternativas para o cálculo convencional do escore utilizado no vestibular. O ajuste do Modelo de Resposta Gradual considerando "errado"como a primeira categoria apresentou resultados absurdos, mas o modelo onde a "não resposta"é a primeira categoria apresentou os melhores resultados entre os modelos propostos, enfatizando que respondentes com maior conhecimento de química tendem a responder um item, correndo o risco de eventualmente errar a resposta, enquanto indivíduos com menor conhecimento tendem a optar pela não resposta ao item.

Palavras-chave: Teoria de Resposta ao item, Modelo de Resposta Gradual, Modelos Dicotômicos, Avaliação, Vestibular, CEBRASPE, Proficiência.

Sumário

1 Introdução

Conhecimento é uma característica difícil de se mensurar em cada indivíduo. A forma mais comum de realizar essa análise seria com a aplicação de provas, porém o resultado de cada indivíduo está atrelado às questões que a prova contém. Logo não podemos comparar o conhecimento de pessoas que realizaram provas diferentes.

Ao longo do tempo, técnicas foram criadas para solucionar este tipo de problema, como é o caso da Teoria de Resposta ao Item (TRI). Basicamente, o que esta metodologia nos sugere é uma relação entre as habilidades ou proficiências de um indivíduo em uma determinada área de conhecimento com a probabilidade deste indivíduo acertar a resposta de uma questão desta área, ou seja, usar o item como elemento central do estudo e não a prova em si. Isso nos permite comparar o desempenho de diferentes populações que realizaram provas com itens similares ou até mesmo comparar indivíduos da mesma população que realizaram provas diferentes.

Uma das maneiras de se ingressar na UnB é através do vestibular, uma prova interdisciplinar elaborada pelo Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (CEBRASPE) para avaliar o conhecimento dos respondentes nas disciplinas . A prova é realizada em 2 dias, possui 4 tipos de questões (A,B,C e D), uma redação e uma parte das questões dedicada á língua estrangeira escolhida pelo respondente (Inglês, Espanhol ou Francês). Veremos mais detalhes sobre esta avaliação posteriormente no estudo.

Neste projeto utilizaremos o Modelo de Resposta Gradual (MRG) da Teoria de Resposta ao item considerando três categorias de alternativas (errado, não resposta, certo e não resposta, errado, certo). Pois este modelo considera que as categorias de resposta de uma questão podem ser ordenadas entre si, ou seja, as categorias mais altas contribuem mais para o escore final do que as categorias mais baixas. Aplicaremos tal modelo diretamente em um banco de dados que contém as respostas de indivíduos que realizaram o vestibular da UnB em 2014, mas apenas nas questões do tipo A e C referentes a disciplina de química, e em seguida comparar os resultados obtidos pelo modelo com os resultados reais. Através dos modelos de TRI, procuramos identificar características da prova, como itens com baixo poder de discriminação.

2 Objetivos

2.1 Objetivo Geral

Avaliar as questões de química do Vestibular da UnB de 2014 via Teoria de Resposta ao Item.

2.2 Objetivos Específicos

- Ajustar modelos de Teoria de Resposta ao Item aos dados em estudo.
- Comparar as notas de química obtidas no vestibular com as notas estimadas pelo Modelo de Resposta Gradual da TRI.
- Verificar possíveis diferenças no modelo MRG proposto com base nas duas alternativas de categorias utilizadas: (errado, não resposta, certo) e (não resposta, errado, certo).
- Identificar questões que tenham pouco poder de discriminação entre indivíduos com níveis distintos de conhecimento.

3 Teoria de Resposta ao Item

Segundo Andrade, Tavares e Valle (2000), a TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade (ou habilidades) do respondente. Essa relação é sempre expressa de tal forma que quanto maior a habilidade, maior a probabilidade de acerto no item.

Os modelos da TRI apresentam a relação entre o traço latente medido pelo instrumento e uma resposta a um determinado item. Sendo que os itens podem ser divididos em dicotômicos, em que há 2 categorias de resposta, e politômicos, com mais de 2 categorias.(DeMars (2010))

Modelos de TRI dependem de três fatores: natureza do Item, número de populações em estudo e a quantidade de traços latentes a serem calculados. Natureza do item pode ser classificada em dicotômica ou não dicotômica, número de populações (uma ou mais de uma população) e quantidade de traços latentes, podendo ser classificado como unidimensional ou multidimensional. Neste trabalho serão estudados itens dicotômicos (de duas categorias de resposta) e politômicos (com mais de duas categorias de resposta)e serão utilizados modelos unidimensionais (que mede apenas um traço latente) para uma população.

3.1 Modelos TRI Para Itens Dicotômicos

3.1.1 Modelo Logístico Unidimensional de 1 Parâmetro

O modelo logístico unidimensional de 1 parâmetro (ML1), também conhecido como modelo de Rasch, expressa a probabilidade do j-ésimo indivíduo acertar o i-ésimo item pela seguinte equação:

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}, i = 1, 2..., I; j = 1, 2..., n. \tag{3.1.1}$$

em que U_{ij} é a resposta do candidato classificada em 1 se certa e 0 se errada, θ_j é a habilidade do candidato, $P(U_{ij} = 1|\theta_j)$ é a função de resposta do item e b_i é o parâmetro de dificuldade do item, b_i representa o ponto na escala da habilidade onde a probabilidade de acertar o item é 0,5.

3.1.2 Modelo Logístico Unidimensional de 2 Parâmetros

O modelo logístico unidimensional de 2 parâmetros (ML2), é similar ao modelo de 1 parâmetro porém é adicionado o parâmetro de discriminação do item (a_i). Este modelo tem sua fórmula dada por :

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}, i = 1, 2..., I; j = 1, 2..., n, \tag{3.1.2}$$

O parâmetro a_i indica o poder de discriminação de um item.

3.1.3 Modelo Logístico Unidimensional de 3 Parâmetros

O modelo logístico unidimensional de 3 parâmetros (ML3) continua sendo similar aos outros dois modelos já vistos mas com a introdução de mais um parâmetro, o parâmetro de acerto ao acaso (c_i) e tem sua equação dada por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}}, i = 1, 2..., I; j = 1, 2..., n, \tag{3.1.3}$$

O parâmetro (c_i) representa a probabilidade de um aluno com baixa habilidade responder corretamente o item e é muitas vezes referido como a probabilidade de acerto ao acaso. Então, quando não é permitido "chutar", (c_i) é igual a 0.

3.2 Curva Característica do Item

A probabilidade de um candidato com habilidade θ acertar o item i pode ser representada de forma gráfica pela Curva Característica do Item (CCI).

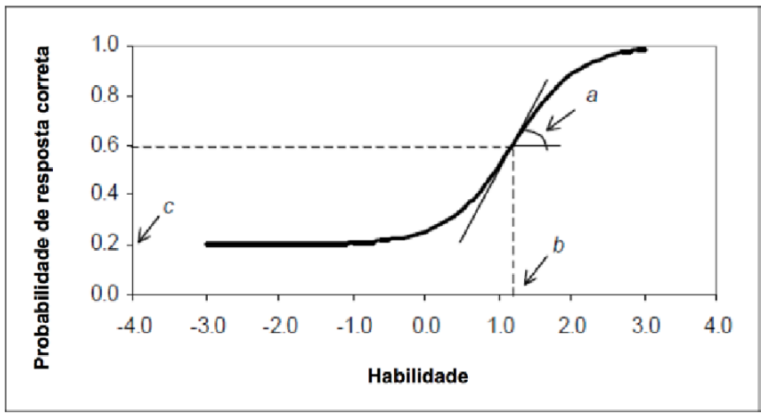

Figura 1: Exemplo de uma Curva Característica do Item

Na Figura 1 acima temos um exemplo de uma CCI para um item qualquer. Nela observamos como a probabilidade de resposta correta de um candidato aumenta a medida que sua habilidade também aumenta. A curva é definida pelos parâmetros a, b e c indicados na figura.

O parâmetro a é o poder de discriminação do i-ésimo item e indica a inclinação da CCI. Quanto maior o valor de a maior será a inclinação da curva, ou seja, maior será a diferença entre as probabilidades de resposta correta de indivíduos com habilidade distintas. b é o parâmetro de dificuldade do item. De forma resumida, quanto maior o valor de b, maior é o grau de dificuldade da questão. E, finalmente, c indica o parâmetro de acerto casual que nada mais é que a probabilidade de um indivíduo com uma habilidade baixa acertar o item. É importante ressaltar que em itens onde o aluno pode chutar, como em itens do tipo A ou C, o parâmetro de acerto ao acaso nunca poderá ser igual a 0.

3.3 Função de Informação do Item

Segundo Fernández et al. (1990), a função de informação dos itens é um poderoso instrumento para análise de itens, possibilitando o conhecimento não só de quanto de informação um item acumula num determinado valor de θ , mas também em que valor de θ o item possui maior quantidade de informação. A função de informação do item é dada por:

$$I_i(\theta) = \frac{\frac{d}{d\theta}[P_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{3.3.1}$$

- $P_i(\theta) = P(U_{ij} = 1 \mid \theta_j)$

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

Adaptando a função para um modelo logístico de 3 parâmetros temos:

$$I_i(\theta) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i} \right]^2 \tag{3.3.2}$$

Esta função agora contém os 3 parâmetros presentes no modelo, e temos que a informação será maior:

1. quando b_i se aproxima de θ;
2. quanto maior for a_i;
3. quanto mais c_i se aproximar de 0.

3.4 Função de Informação do Teste

A função de Informação do Teste nada mais é que a soma das informações dos itens, dada por:

$$I_i(\theta) = \sum_{i=1}^{I} I_i(\theta) \tag{3.4.1}$$

A função de informação está relacionado com o erro padrão de estimação da seguinte forma:

$$EP_i(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3.4.2}$$

3.5 Modelo de Resposta Gradual

O modelo de resposta gradual (MRG) proposto por Samejima (1969) assume uma ordenação nas categorias de respostas do itens, onde as categorias de resposta mais baixas indicam proficiência mais baixa e as mais altas níveis mais altos de proficiência.

Este modelo tenta obter mais informação das respostas dos indivíduos do que simplesmente se eles forneceram respostas corretas ou incorretas.

Suponha as categorias de um item i são arranjadas em ordem crescente e denotados por $k = 0, 1, \ldots, m_i$, onde $(m_i + 1)$ é o número de categorias do i-ésimo item. A probabilidade de um indivíduo j escolher uma categoria particular k ou outra mais alta do item i é dada por:

$$P^{+}_{i,k}(\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_{i,k})}} \tag{3.5.1}$$

com $i = 1, 2, ..., I$, $j = 1, 2, ...n$, e $k = 1, 2, ..., m_i$, onde:

- $P^{+}_{i,k}(\theta_j)$ é a probabilidade de um indivíduo j escolher a categoria k ou outra mais alta do item i;
- $b_{i,k}$ é o parâmetro de dificuldade da k-ésima categoria do item i;
- θ_j representa a habilidade, proficiência (traço latente) do j-ésimo indivíduo;
- a_i é o parâmetro de discriminação do item i, com valor proporcional à inclinação da CCI no ponto;

O parâmetro de discriminação a varia a cada item, mas é constante dentro dos itens. Essa restrição de igual inclinação em cada categoria tem a finalidade de evitar probabilidades negativas.

Para o parâmetro de dificuldade $b_{i,k}$, por definição, temos que:

$$b_{i,1} \leq b_{i,2} \leq ... \leq b_{i,m_i} \tag{3.5.2}$$

A probabilidade do j-ésimo indivíduo receber um escore k para o i-ésimo item, é dada pela expressão

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

Também podemos definir

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

e

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

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

e

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

Então temos que:

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

Note que em um item com (m_i+1) categorias, m_i valores de dificuldade necessitam ser estimados, além do parâmetro de discriminação do item. Assim, para cada item, o número de parâmetros a ser estimado será dado pelo seu número de categorias de resposta.

4 Metodologia

Traços latentes são características de indivíduos que não podem ser mensuradas diretamente. Ao longo do tempo muitos métodos das mais diversas áreas de conhecimento foram desenvolvidos para medir estas características de uma forma adequada, uma delas sendo a Teoria Clássica da Medida (TCM), que utiliza o escore em uma prova como sua referência de medida da proficiência, No entanto, há limitações neste método, pois como os resultados sempre dependem do conjunto de questões utilizadas na prova seria inviável comparar o escore de indivíduos que realizaram provas diferentes.

Posteriormente a Teoria de Resposta ao Item (TRI) foi criada com o intuito de resolver os problemas da TCM que seria a dependência entre o escore obtido e a prova aplicada. Segundo Andrade, Tavares e Valle (2000), A TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade (ou habilidades) do respondente. Essa relação é sempre expressa de tal forma que quanto maior a habilidade, maior a probabilidade de acerto no item. Em outras palavras, um item mede um determinado conhecimento independente de quem está respondendo, com as repostas aos itens podendo ser classificadas como dicotômicas ("sim"e "não"ou "certo"e "errado") ou politômicas com mais de duas categorias.

A principal diferença entre o escore obtido no vestibular pelo método tradicional e o escore obtido via TRI é que pelo cálculo convencional a pontuação obtido pelo acerto de uma questão é sempre igual. Já o escore obtido via TRI, um candidato ganha mais pontos se acertar um item difícil do que um item fácil, fazendo com que o escore obtido via TRI mensure melhor a proficiência do candidato do que o escore convencional padrão.

Como o objetivo desse estudo é avaliar as questões tipo A e C de química do vestibular da UnB, utilizaremos o modelo de resposta gradual (MRG) de Samejima (1969). O MRG é um modelo que, tenta obter mais informações sobre o respondente que simplesmente se ele deu a resposta certa ou errada, assumindo que as categorias de resposta de um determinado item podem ser ordenadas entre si, ou seja, categorias mais altas irão fornecer escores maiores que categorias mais baixas.

Ajustaremos o modelo de resposta gradual da Teoria de Resposta ao Item às questões dos tipos A e C de química com duas alternativas de categorias: (errado, não resposta, certo) e (não resposta, errado, certo). A primeira considerando a categoria "não resposta"como uma resposta intermediária entre certo e errado já que esta não resulta em uma perda ou ganho de pontos, e a segunda considerando a categoria "não resposta"como sendo a categoria mais baixa e em seguida comparar os escores obtidos pela estimação com os escores reais dos respondentes.

4.1 Material

Como citado anteriormente, o material analisado neste estudo é proveniente das respostas de indivíduos que realizaram o vestibular da UnB do segundo semestre de 2014 disponibilizado pelo CEBRASPE. A avaliação foi divida em 3 partes e realizada durante 2 dias. Em cada dia os candidatos possuíam 5 horas para realizar a avaliação. No primeiro dia, foi realizada a parte 1, referente às provas de Língua Estrangeira (Inglês, Espanhol e Francês) com 30 questões,a parte 2 referente às disciplinas de "Humanas"(Português, Literatura, Geografia, História, Artes, Filosofia e Sociologia) com 120 questões e finalmente os candidatos deveriam realizar uma redação. Já no segundo dia os candidatos deveriam responder a 150 questões referente às disciplinas de exatas (Biologia, Física, Química e Matemática). Dentro das 300 questões apresentadas na avaliação, existem 4 tipos diferentes de itens: itens do tipo A, B, C e D. Neste estudo iremos apenas considerar itens do tipo A (certo ou errado) e do tipo C (múltipla escolha dentre 4 alternativas). Vale ressaltar que nas questões de tipo A o candidato ganha 1 ponto caso acerte o item, perde 1 um ponto caso erre e não ganha nenhum ponto quando não responde. Já para as questões de tipo C o respondente ganha 2 pontos quando acerta, perde 0.667 pontos quando erra e novamente ganha zero quando não responde. Porém haviam 3 tipos de provas (Tipo I, II e III) que se diferenciam apenas na ordem de algumas questões. Além disso, na parte 1 (Língua Estrangeira) o candidato poderia optar por realizar a prova em qualquer uma das três línguas disponíveis. Desta forma dentro de cada tipo de prova haviam 3 possíveis formas de resolução, totalizando em um total de 9 tipos de provas distintas.

Depois dessa constatação, neste presente estudo iremos considerar apenas as 38 questões do tipo A ou C da disciplina de química respondidas por 7.232 candidatos que realizaram o caderno de prova de tipo I no segundo dia do vestibular.

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 2: Características da prova do vestibular da UnB 2014

5 Resultados

Neste capítulo, serão apresentados os resultados obtidos a partir das análises do banco de dados. Os resultados foram obtidos por análises feitas no software R

5.1 Análise Descritiva

5.1.1 Questões Tipo A

A tabela 1 mostra a distribuição de frequência das respostas para as questões do tipo A, vemos que a maioria das respostas dos candidatos foi correta representando 38,05% das respostas, em seguida temos que 36,8% das respostas dos candidatos foram não respostas, uma frequência muito similar a de respostas corretas. Por fim, temos que 25,14% das repostas dos candidatos foi incorreta.

Tabela 1: Frequência das Respostas à Questões do Tipo A

Resposta	Frequência Absoluta	Frequência Relativa
Certa	93573	38,05%
Não Resposta	90493	36,8%
Errada	61822	25,14%
Total	245888	100%

As respostas também podem ser representadas graficamente pelo gráfico abaixo.

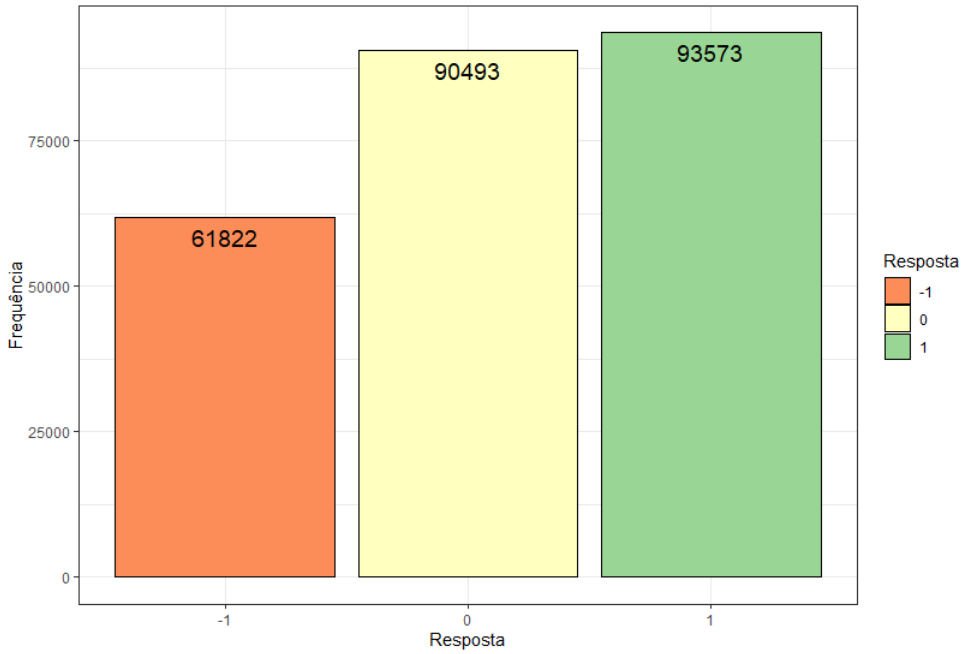

Figura 3: Gráfico de barras da frequência das respostas

5.1.2 Questões Tipo C

Para as questões do tipo C, a tabela abaixo nos apresenta a frequência para cada categoria de resposta. A categoria correta representando 40,63% das respostas e a categoria não resposta com apenas 20,71% das respostas. E por fim a categoria errada, agora correspondendo a 38,65% das respostas. Esta diferença de distribuição de respostas quando comparada às questões de tipo A se dá justamente pelo diferente sistema de pontuação das questões de tipo C, onde a pontuação de acertar uma questão é três vezes maior que a penalidade de errar, fazendo que mesmo contendo 4 possíveis alternativas de resposta, candidatos tentem arriscar o acerto neste tipo de questão.

Tabela 2: Frequência das Respostas à Questões do Tipo C

Resposta	Frequência Absoluta	Frequência Relativa
Certa	11755	40,63%
Não Resposta	5992	20,71%
Errada	11181	38,65%
Total	28.928	100%

Também podemos analisar as respostas pelo gráfico abaixo.

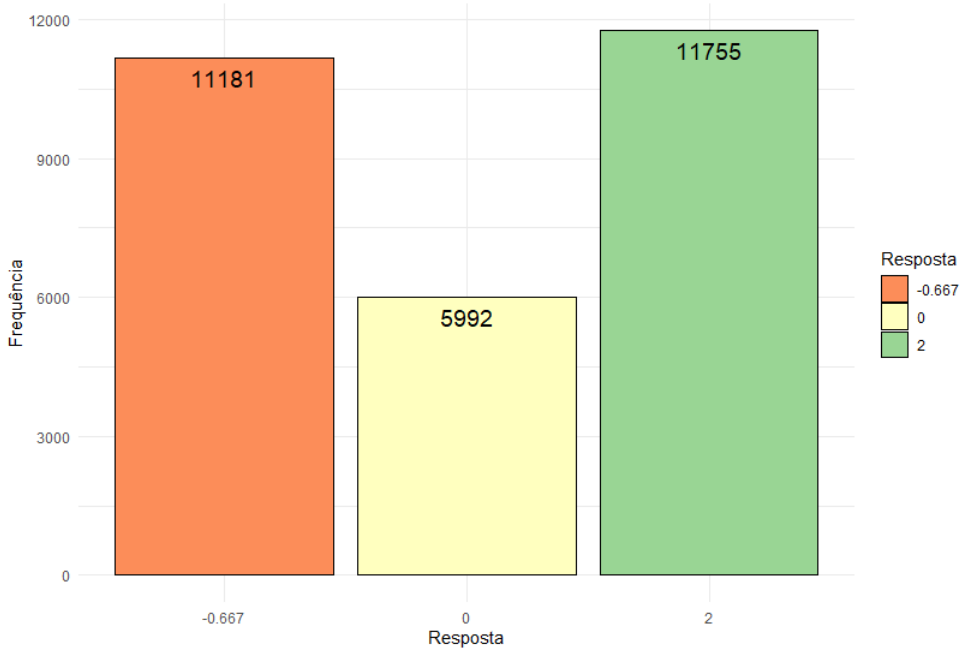

Figura 4: Gráfico de barras da frequência das respostas

5.1.3 Escores

Iremos calcular a nota obtida por cada candidato nas questões de química considerando os escores já citados. Se um candidato acertar todas as 34 questões do tipo A e todas as 4 do tipo C, ele ira obter o escore máximo de 42 pontos. Por outro lado, se errar todas as questões, ele irá obter um total de -36,66 pontos.

A tabela abaixo apresenta algumas estatísticas dos escores obtidos pelos candidatos. Podemos observar que o menor escore é -20. Já o maior é de 38, nota muito próxima do escore máximo. Percebemos um valor muito baixo tanto para média quanto para mediana de respectivamente, 5 e 6.61, o que indica de uma forma geral um mau desempenho dos candidatos nas questões de química.

Tabela 3: Estatísticas dos Escores dos Candidatos

Mínimo	-20
1º Quartil	0.33
Mediana	5
Média	6.61
3º Quartil	11.33
Máximo	38

Também podemos analisar as notas de forma gráfica pelo histograma abaixo, Como já observado na tabela, percebemos uma grande concentração de notas entre 0 e 10. Vemos, no gráfico, que os escores têm uma distribuição assimétrica à direita.

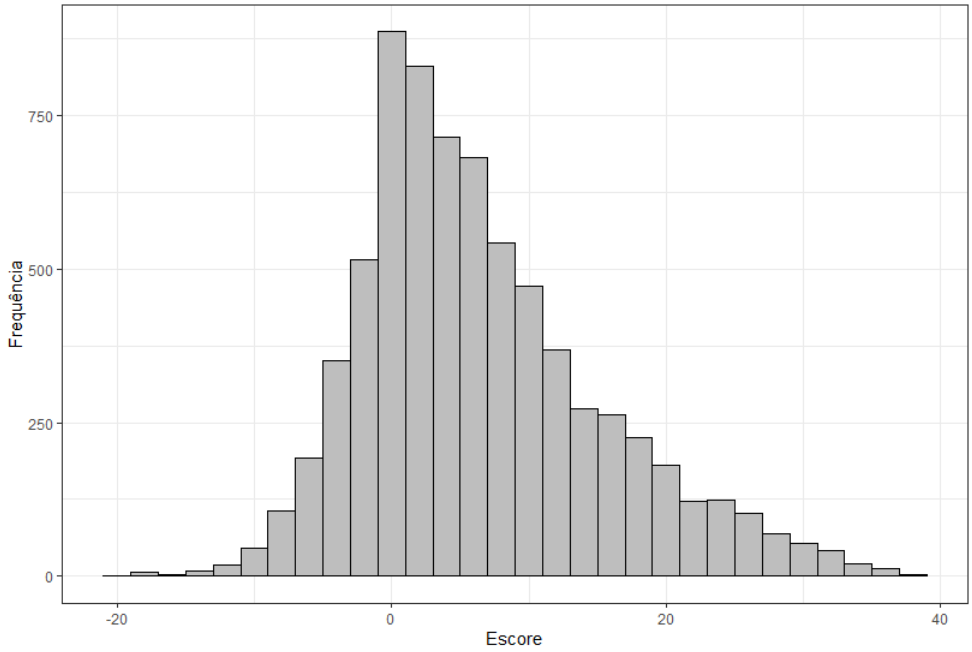

Figura 5: Histograma da Distribuição do Escore dos Candidatos

5.1.4 Frequência de Respostas dos Itens

Pela Figura 6 verifica-se a frequência de resposta de cada item. Observamos a notável diferença de frequência de resposta entre os itens. Os itens como o 5, 8, 12, 14, 21, 24, 26 e 32 apresentam uma grande frequência de respostas corretas, o que pode indicar que esse sejam itens com um baixo grau de dificuldade. Em contrapartida os itens 4, 19, 35 e 38 apresentam uma alta frequência de respostas incorretas, possivelmente por serem itens com um alto grau de dificuldade. Já os itens 6, 9, 25 e 30 tem uma grande quantidade de não resposta.

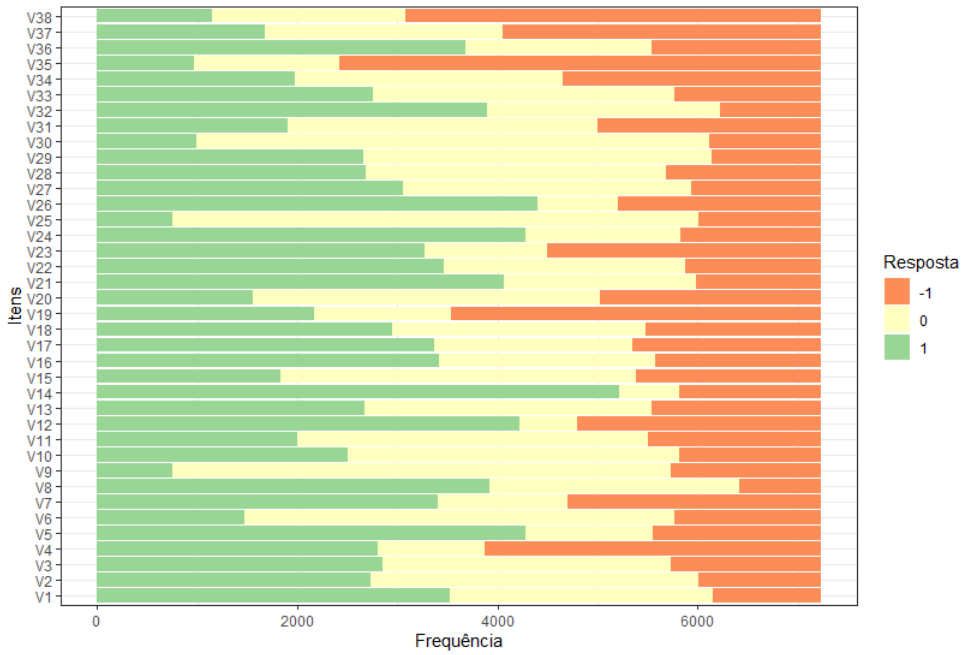

Figura 6: Gráfico da frequência de acerto, não resposta e erro de cada questão.

5.1.5 Proporção de Acerto nos Itens

Partindo agora para uma verificação gráfica da proporção de acerto nas questões, pela Figura 7 constatamos que os itens com o maior grau de dificuldade, ou seja, com a menor proporção de acerto são os itens 25, 9, 20, 30 e 38 com as taxas de acerto de 10,48%, 10,49%, 13,3%, 13,7% e 15,9% respectivamente. Por outro lado, os itens com o maior grau de acerto são os itens 14, 26, 24, 5 e 12 com taxas de acerto de 72%, 61%, 59,25%, 59,22% e 58,4%. Vale ressaltar que os itens 26, 38 e 12 são itens do tipo C.

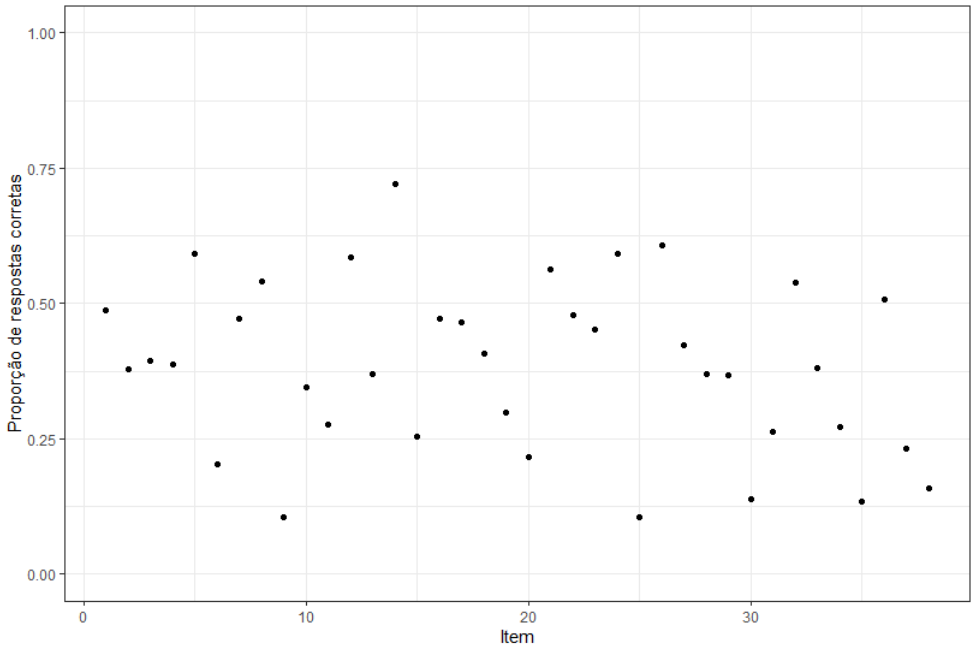

Figura 7: Plot da taxa de proporção de acerto de cada item

5.1.6 Proporção de Acerto de Acordo com O Escore do Candidato

A proporção de acerto de cada item de acordo com o escore dicotomizado de cada candidato seria equivalente a uma estimativa não paramétrica da CCI de cada item. A figura 8 mostra este gráfico para cada uma das 38 questões. Itens com um alto grau de dificuldade apresentam baixa proporção de respostas corretas mesmo para candidatos com escores dicotomizados altos, o que pode ser observado nos itens 6, 9, 25, 30 e 35. O escore dicomotizado é o escore de cada candidato considerando 0 quando erra ou não responde o item e 1 quando responde corretamente o item.

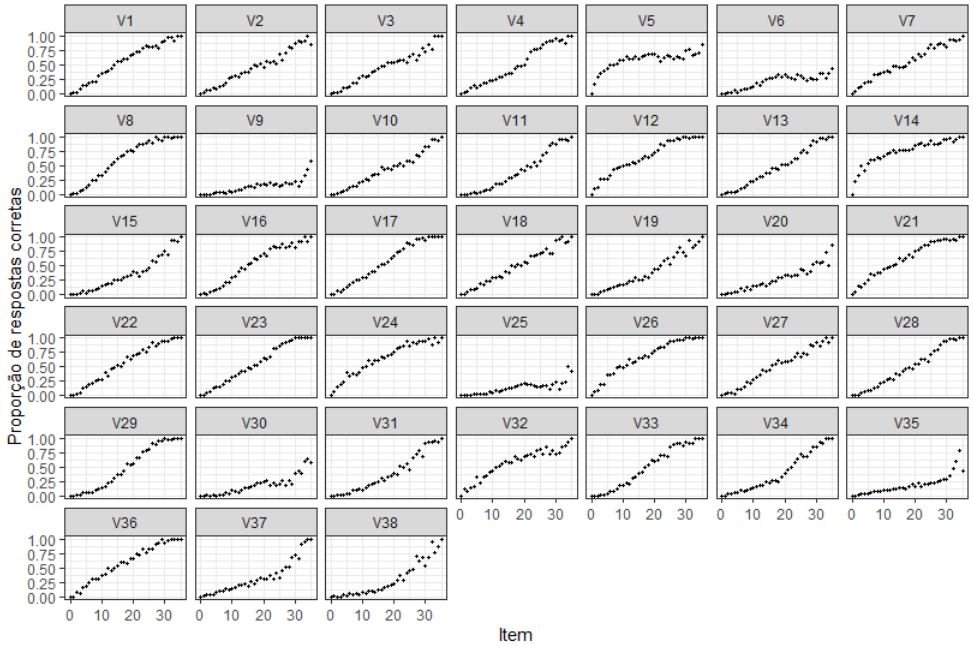

Figura 8: Gráfico da taxa de proporção de acerto de cada item

5.2 Modelo Logístico de 3 Parâmetros

Partindo agora para a resolução do objetivo principal deste estudo, avaliar as questões de química dos candidatos, utilizaremos primeiramente o modelo logístico dicotômico de 3 parâmetros (ML3), considerando apenas duas categorias de resposta. Neste caso, o modelo considera não resposta e resposta errada como uma única categoria. A tabela 4, apresenta medidas padrões dos parâmetros e da proficiência estimada.

Com base na tabela, podemos observar os valores não negativos para o parâmetro de discriminação (a_i), variando entre 0.1615 e 2.1915, indicando haver itens com baixo poder de discriminaçao entre indivíduos com diferentes níveis de proficiência. Para o parâmetro de dificuldade (b_i) constatamos valores variando entre -2.3170 e 5.0661. 5 é um valor muito alto para o parâmetro de dificuldade, já que está na escala da distribuição $N(0, 1)$. Já para o parâmetro de acerto casual (c_i) percebemos que a grande maioria dos itens apresenta valor igual a (ou próximo de) zero pois como citado anteriormente o vestibular penaliza por cada item respondido errado, fazendo com que os candidatos não arrisquem acertar ao acaso uma questão. O item com o maior valor de parâmetro de acerto ao acaso é a questão 12 que é uma questão do tipo C que penaliza menos o candidato caso ele erre. Finalmente a proficiência estimada (θ_j) variando entre -2.1986 e 3.0074 nos mostra a grande diferença de conhecimento entre os candidatos o que é de se esperar de uma prova aplicada a uma grande e diversa população como o vestibular da UnB.

Medidas	Parâmetro de Discriminação (a_i)	Parâmetro de Dificuldade (b_i)	Parâmetro de Acerto Casual (c_i)	Proficiência Estimada (θ_j)
Mínimo	0.1615	-2.3170	0	-2.1986
1 Quartil	0.7297	0.2516	0	-0.5895
Mediana	1.0367	0.6178	0	0.02165
Média	1.0836	0.9850	0.0506	0.0023
3 Quartil	1.4730	1.1751	0.0644	0.5440
Máximo	2.1915	5.0661	0.3621	3.0074

Tabela 4: Tabela com medidas padrões dos parâmetros e proficiência

A Figura 9 abaixo mostra a curva característica de todos os itens analisados. A maioria dos itens apresenta a curva bastante similar, porém, é fácil notar que itens com baixo poder de discriminação, como os itens 25, 9, 5, 35 e 30, apresentam curvas com um baixo grau de inclinação significativamente diferentes das demais. Itens com um parâmetro de acerto ao acaso c_i significativamente maior que 0, como os itens 12, 26, 21,

24 e 7, apresentam o início da curva em pontos maiores que os demais.

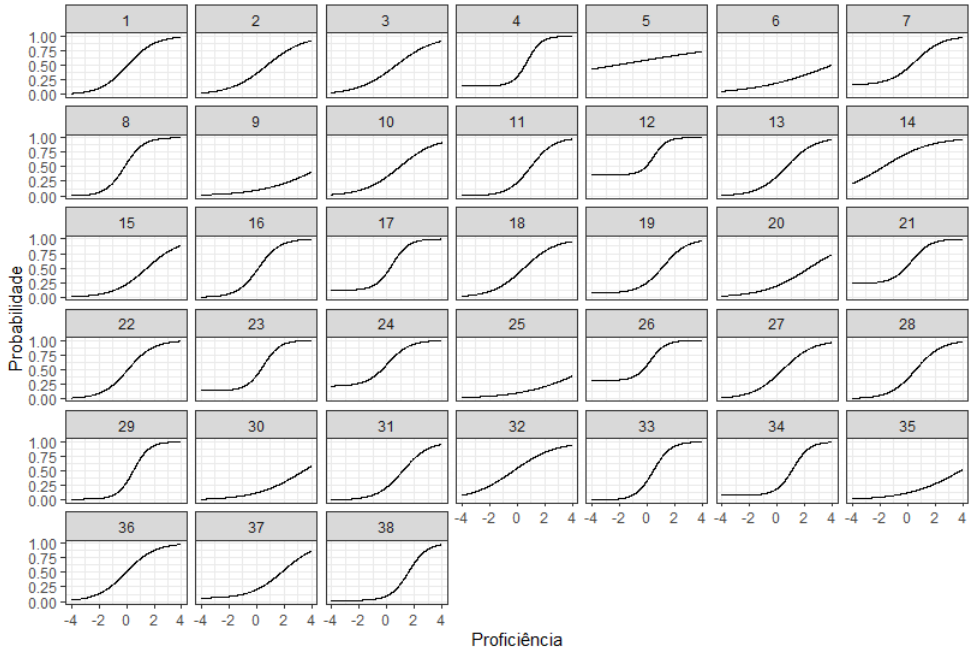

Figura 9: Gráfico da CCI de todos os itens considerando o ML3

A Figura 10 abaixo mostra a proporção de respostas corretas e o parâmetro de dificuldade (b_i) de cada item. É fácil notar que, à medida que o parâmetro de dificuldade diminui, a proporção de resposta correta para cada item aumenta, o que é esperado pois item mais fáceis terão mais respostas corretas que item mais difíceis.

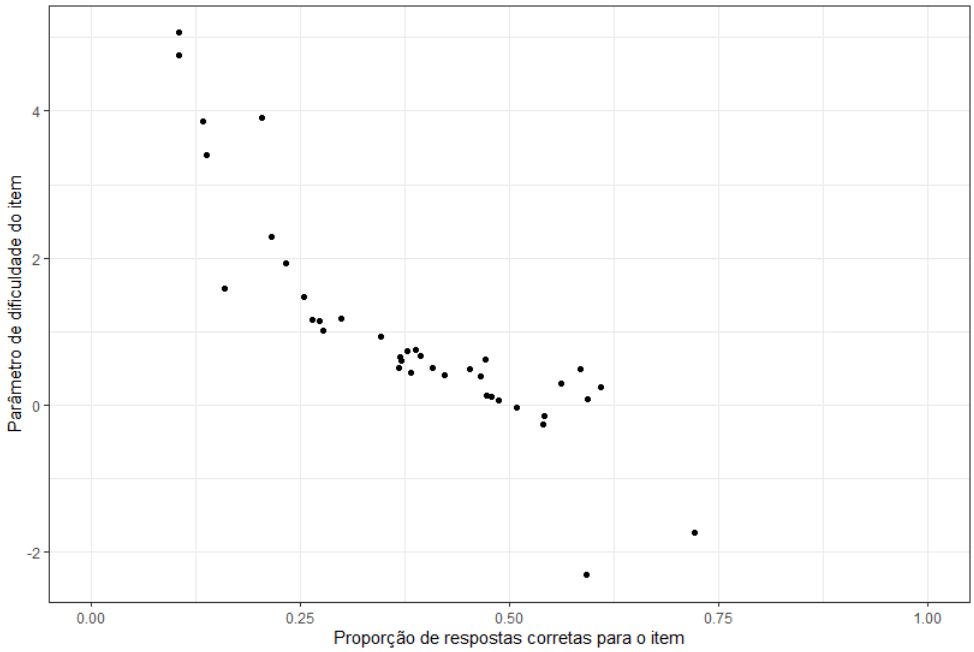

Figura 10: Gráfico da taxa de proporção de acerto versus o parâmetro de dificuldade de cada item

Em seguida a figura 11 apresenta a proficiência estimada de cada candidato (θ_j) versus o escore convencional do vestibular padronizado, evidenciamos uma forte correlação entre o escore e a proficiência, com o coeficiente de correlação de Pearson de 0.7878 indicando uma correlação muito forte, o que mostra que o modelo logístico de 3 parâmetros estima a proficiência de forma aceitável.

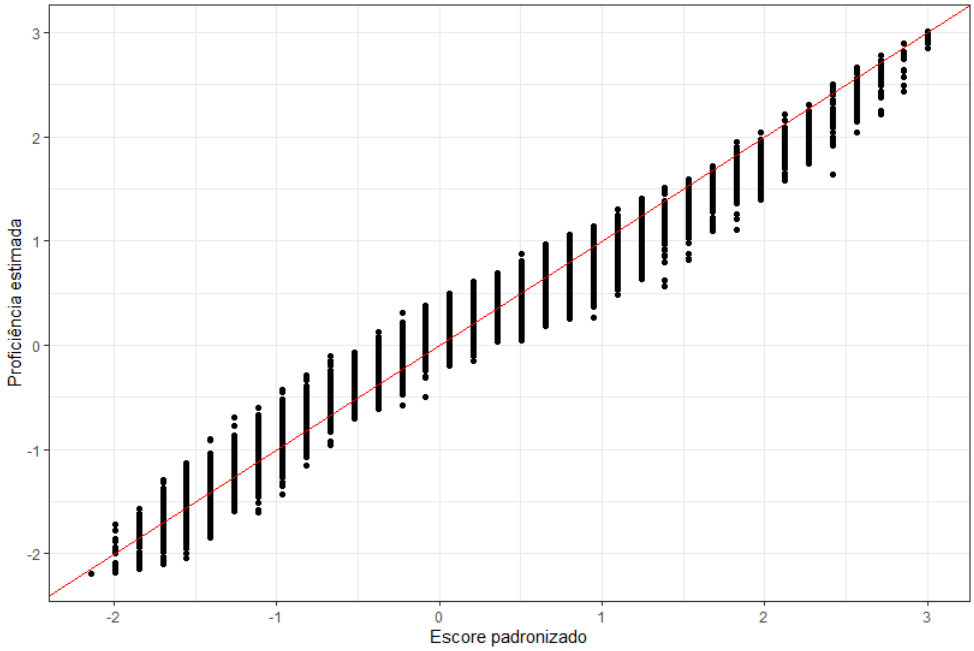

Figura 11: Plot do escore padronizado versus proficiência estimada para o modelo logístico de 3 parâmetros

Já a figura 12 nos mostra o gráfico do número de respostas corretas pelo parâmetro de acerto ao acaso (c_i) de cada item. Como já relatado anteriormente vemos que a grande maioria dos itens apresenta parâmetro de acerto ao acaso próximo de zero e pelo gráfico também vemos que itens que apresentam um valor de acerto ao acaso significativo tem um grande número de respostas corretas. Isto se dá porque estes itens apresentam parâmetro de dificuldade mais baixo.

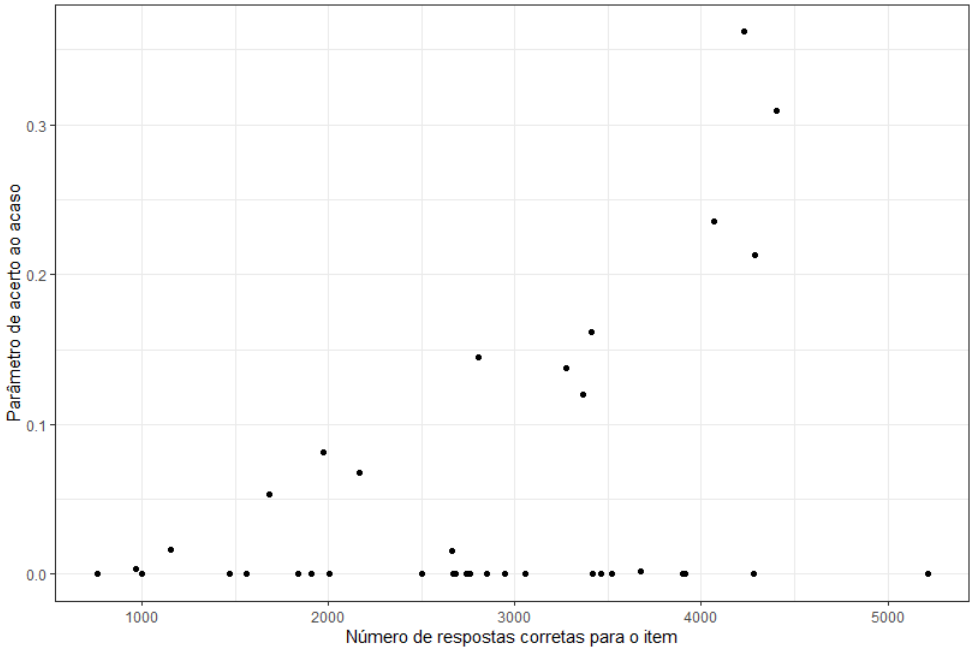

Figura 12: Plot do número de respostas corretas versus o valor do parâmetro de acerto ao acaso de cada item

5.3 Modelo Logístico de 2 Parâmetros

O Modelo logístico dicotômico de 2 parâmetros (ML2) é muito similar ao ML3 porém não contém o parâmetro de acerto ao acaso (c_i) em sua fórmula. Iremos verificar o ajuste deste modelo pois como visto no modelo anterior, a grande maioria dos itens apresenta o parâmetro de acerto ao acaso igual a zero, o que questiona a necessidade deste parâmetro. A tabela 5 apresenta medidas resumo dos parâmetros estimados e da proficiência estimada dos candidatos.

Fazendo uma análise da tabela, vemos que novamente o parâmetro de discriminação apresenta valores não negativos variando entre 0.1574 e 1.6482 lembrando que valores negativos de a_i não são esperados neste modelo. Para o parâmetro de dificuldade visualizamos valores variando entre -2.3865 e 5.0922 mostrando novamente um valor máximo muito alto e a grande diferença de dificuldade entre os itens como esperado. E por fim, uma grande variação na proficiência estimada também indicando a disparidade

de conhecimento dos candidatos.

Medidas	Parâmetro de Discriminação (a_i)	Parâmetro de Dificuldade (b_i)	Proficiência Estimada (θ_j)
Mínimo	0.1574	-2.3865	-2.4367
1 Quartil	0.6951	0.0686	-0.5856
Mediana	0.9441	0.4908	-0.0066
Média	0.9203	0.8687	0.0001
3 Quartil	1.1326	1.1366	0.5326
Máximo	1.6482	5.0922	3.0669

Tabela 5: Tabela com medidas padrões dos parâmetros e proficiência

A Imagem 13 abaixo mostra o gráfico das CCIs de todos os itens. Notamos que as curvas neste modelos são bem similares aos do ML3. No entanto, por este modelo não possuir o parâmetro de acerto ao acaso c_i, a maioria das curvas se inicia em valores bem próximos de zero. Exceto as curvas dos itens 5 e 14 que se iniciam em valores significativamente acima de zero pelo baixo valor do parâmetro de dificuldade b_i apresentado por estes dois itens, de -2.38 e -1.71 respectivamente.

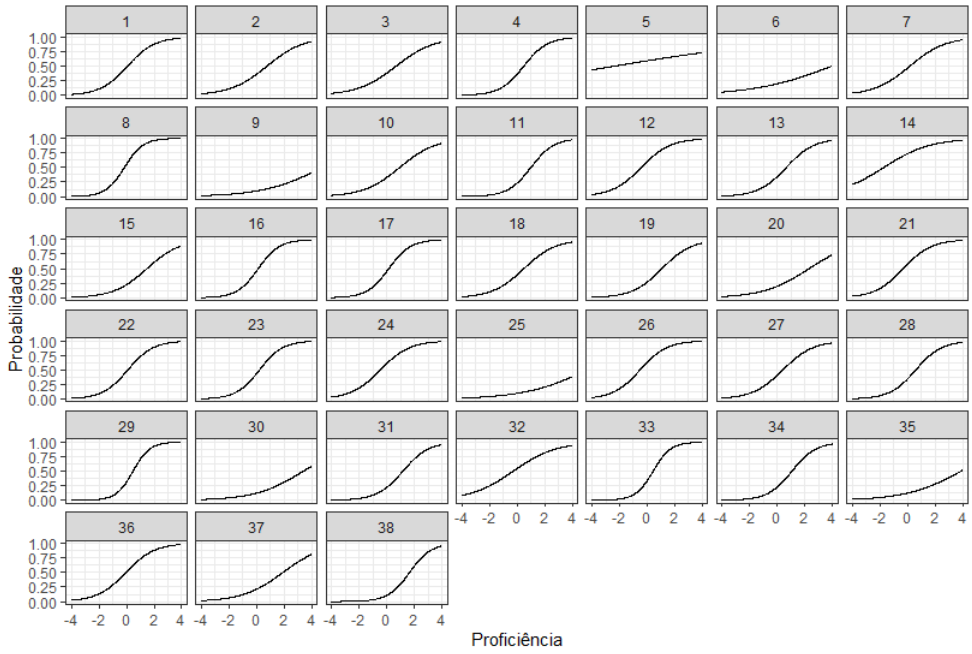

Figura 13: Plot da CCI de todos os itens considerando o ML2

É interessante notar que, mesmo sem o parâmetro de acerto ao acaso c_i, as estimativas obtidas nos parâmetros do ML2 ficam bem similares as obtidas no ML3. Os gráficos abaixo tem o intuito de evidenciar a similaridade destes modelos.

A Figura 14 nos mostra um gráfico de comparação entre os parâmetros de discriminação do ML2 e do ML3. Vemos que em sua maioria os itens têm praticamente o mesmo valor do parâmetro, exceto por alguns itens que apresentam valor de a_i maior em ML3 do que em ML2.

Já a Figura 15 mostra a comparação do parâmetro de dificuldade b_i dos modelos ML2 e ML3. Novamente, a maioria dos itens apresentam resultado muito similar e alguns poucos itens apresentam valor de b_i um pouco maior no ML3 do que no ML2.

E finalmente a figura 16 compara a proficiência estimada dos 7232 candidatos calculado pelo ML2 e pelo ML3, percebemos diferenças nas proficiências de valores mais baixos, variando de -2 a 0, onde o ML2 apresenta valores de θ_j maior que o ML3, porém, a partir de 0 as proficiências seguem bem semelhantes, mostrando que os modelos apresentam resultados equivalentes para candidatos com proficiência mais alta.

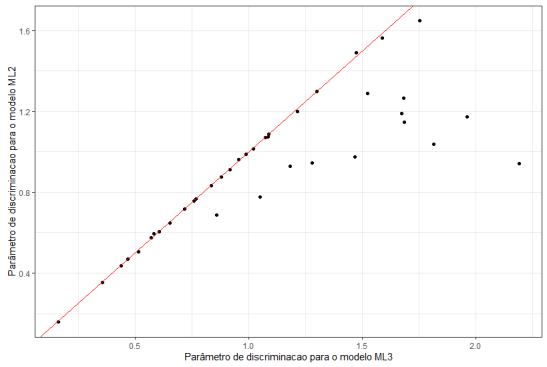

Figura 14: Plot do parâmetro de discriminação do ML2 versus o do ML3

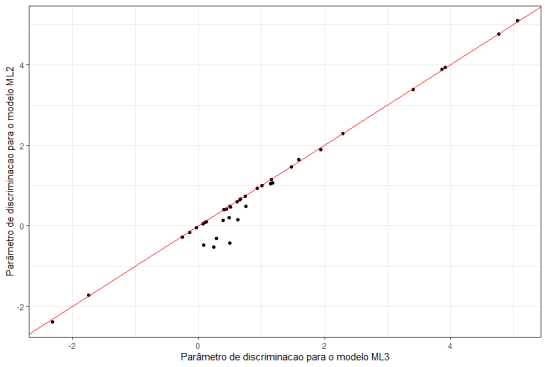

Figura 15: Plot do parâmetro de dificuldade do ML2 versus o do ML3

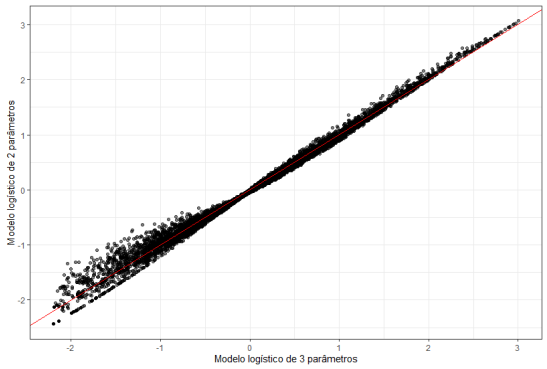

Figura 16: Plot da proficiência estimada do ML2 versus o do ML3

Visto tais semelhanças, agora vamos buscar descobrir o que causou diferença nos parâmetros de determinados itens.

A figura 17 abaixo mostra um gráfico onde no eixo y temos o valor do parâmetro de acerto ao acaso c_i calculado no ML3 de cada item e no eixo x temos a diferença do valor

do parâmetro b_i entre o ML2 e ML3. Constatamos que itens com o valor de c_i próximo de zero, em sua grande maioria não apresentam diferença no valor de b_i, porém a medida que c_i aumenta, a diferença entre o parâmetro de dificuldade dos modelos também aumenta. Expondo que itens com parâmetro de acerto ao acaso alto, como os itens 4, 7, 12, 17, 21, 23, 24, 26, 34 e 38, tendem a apresentar grande diferença no parâmetro de dificuldade nos modelos ML3 e ML2.

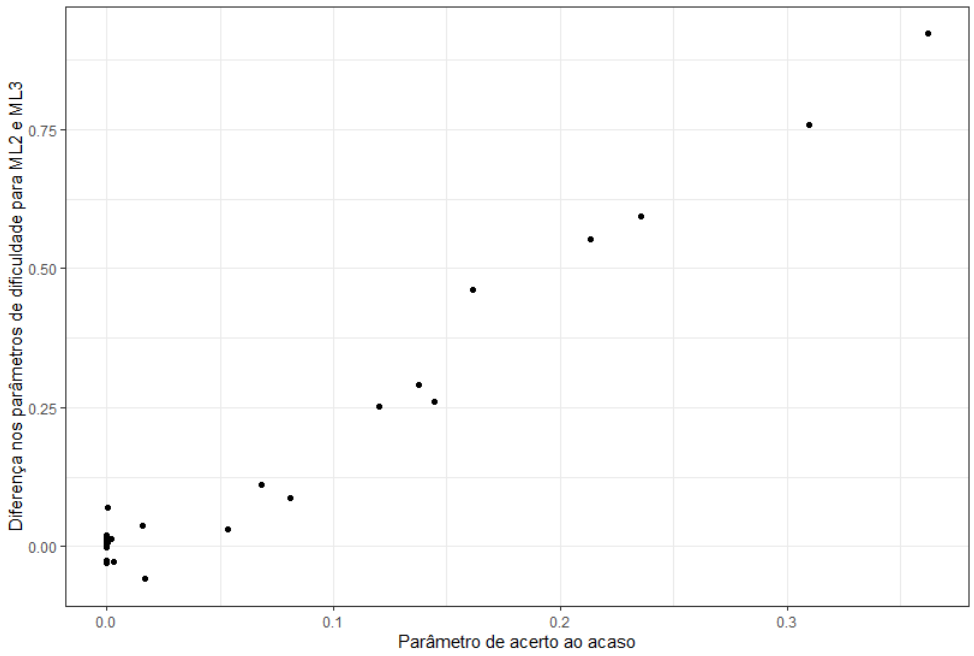

Figura 17: Plot da diferença de b_i entre os modelos versus o valor de c_i de cada item

As figuras 18, 19, 20 e 21 abaixo mostram 4 gráficos em cada. Os boxplots do escore padronizado, proficiência estimada via ML2, proficiência estimada via ML3 e escore convencional do vestibular para os candidatos separados pelo tipo de resposta (errada, não resposta, correta) dada à questão, para as questões 2, 7, 12 e 15 que foram escolhidas ao acaso. Podemos ver que em todas as imagens, nos três primeiros gráficos, respondentes que erram a questão tendem a ter maior estimativa da proficiência na prova do que candidatos que não respondem. Já no último gráfico, do escore convencional do vestibular, notamos que os candidatos que respondem errado e os que não respondem apresentam distribuição similar para a proficiência, mesmo com os candidatos que respondem errado tendendo a apresentar proficiência maior do que aqueles que não respondem.

Isto nos mostra que uma possível ordenação para considerar um modelo politômico de respostas categorizadas seria a ordenação das categorias 1=não respostas, 2=errado e 3=certo.

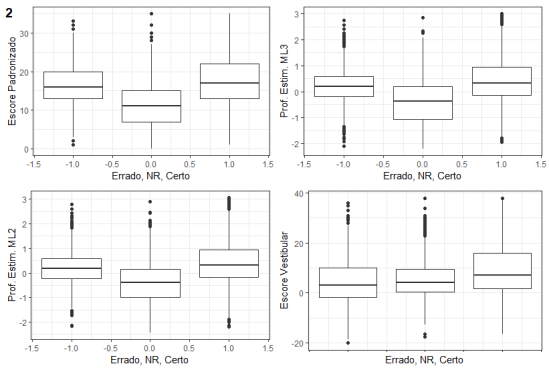

Figura 18: Gráficos da questão 2

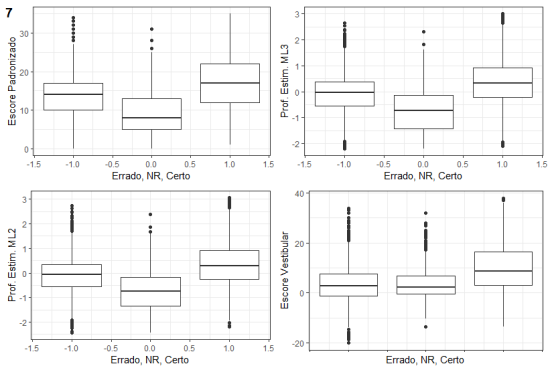

Figura 19: Gráficos da questão 7

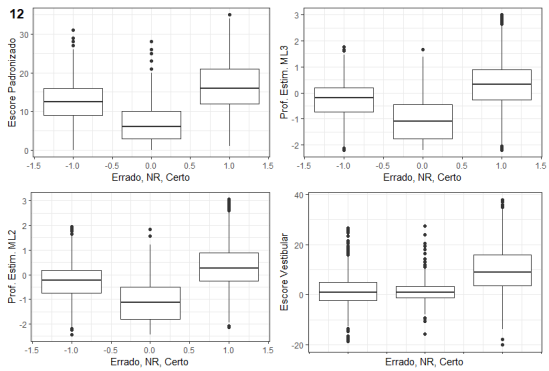

Figura 20: Gráficos da questão 12

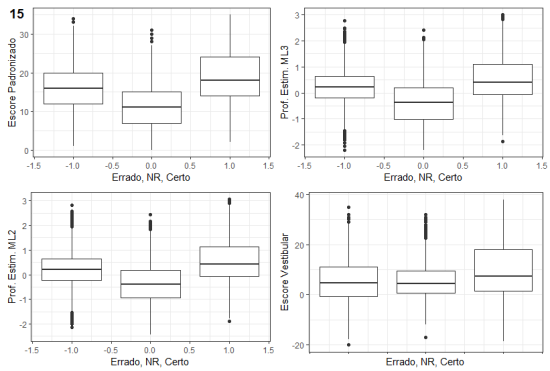

Figura 21: Gráficos da questão 15

5.4 Modelo de Resposta Gradual 1

Buscando encontrar um modelo com desempenho melhor que os modelos dicotômicos já apresentados, iremos utilizar agora o modelo de resposta gradual (MRG), que irá analisar o efeito da "não resposta"na estimação do modelo. Primeiramente iremos utilizar uma ordenação considerando "errado"como a pior categoria, ou seja, 1=errado, 2=não resposta e 3=certo. A tabela 6 abaixo apresenta as medidas resumo dos parâmetros do modelo e da proficiência estimada dos candidatos

Analisando a tabela apresentada abaixo, constatamos que o parâmetro de discriminação (a_i) apresenta valor mínimo de -0.2886, sendo que valores negativos de a_i não são esperados no MRG. Para os parâmetros de dificuldade b_{i1} e b_{i2} temos valores mínimos absurdamente baixos de -299.1026 e -813.0362 respectivamente. Para as proficiências estimadas θ_j a tabela nos mostra valores coerentes. Estes fatos nos mostram de imediato que o modelo pode não estar devidamente ajustado.

Medidas	Parâmetro de Discriminação (a_i)	Parâmetro de Dificuldade (b_{i1})	Parâmetro de Dificuldade (b_{i2})	Proficiência Estimada (θ_j)
Mínimo	-0.2886	-299.1026	-813.0362	-2.4227
1 Quartil	0.2037	-2.9036	-0.3906	-0.6093
Mediana	0.6205	-1.8217	0.2784	-0.2113
Média	0.5625	-7.6703	-23.4815	0.0003
3 Quartil	0.9138	-0.7764	1.4660	0.4667
Máximo	1.5190	33.2602	6.5294	3.1541

Tabela 6: Tabela com medidas padrões dos parâmetros e proficiência do MRG1

A figura 22 abaixo mostra o gráfico das CCIs de todos os itens sendo a linha vermelha desenvolvida com o parâmetro de dificuldade b_{i1} e a linha azul elaborada com b_{i2}. Observamos curvas decrescentes para os itens 30, 5, 6, 25, 9 e 35 pois estes apresentam valor negativo para o parâmetro de discriminação a_i. Já os itens 2, 3, 10, 11, 13, 14, 15, 20, 27, 32 e 37 apresentam valor do parâmetro de discriminação positivo porém baixo, causando que possuam curvas com um baixo nível de inclinação.

As curvas definidas utilizando b_{i1} sempre apresentam maiores probabilidades de acerto que as curvas em que se usa b_{i2}, em itens com baixo parâmetro de discriminação esta diferença se torna ainda mais absurda.

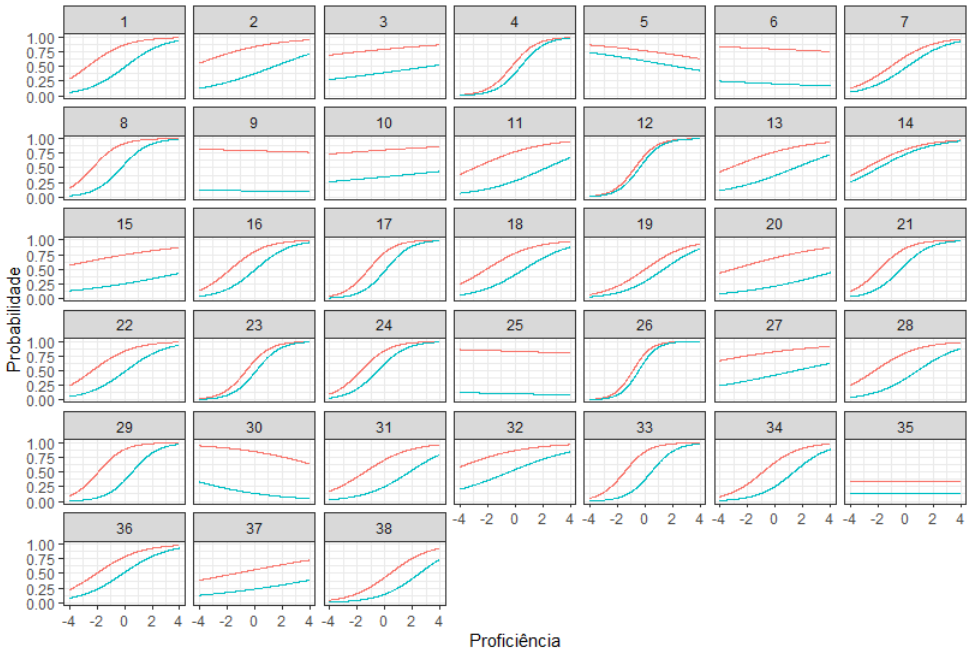

Figura 22: Gráfico da CCI de todos os itens considerando o MRG1

A figura 23 mostra o gráfico do parâmetro de discriminação de cada item. Notamos que os itens 30, 5, 6, 25, 9, e 35 apresentam valores negativos. Isto faz com que estes itens apresentem suas CCIs decrescentes, o que não faz sentido e como já dito anteriormente não é esperado neste tipo de modelo.

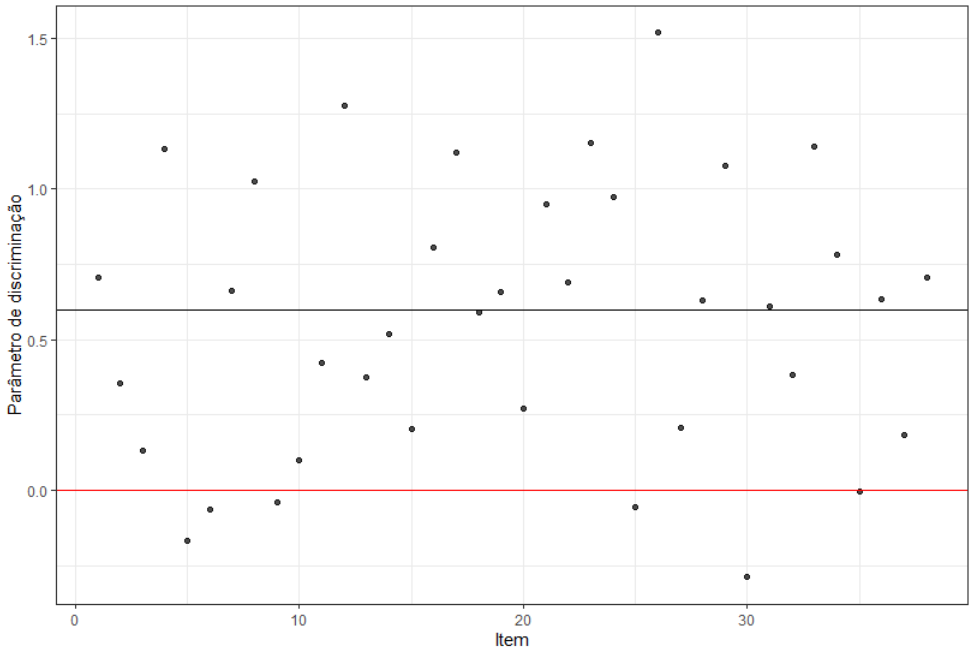

Figura 23: Plot do valor do parâmetro de discriminação a_i de cada um dos itens.

Já a figura 24 nos mostra os valores do parâmetro de dificuldade b_{i1} em vermelho e b_{i2} em azul. Os itens 35, 9, 25, e 6 apresentam os valores destes parâmetros completamente absurdos, com $b_{i1} = -299.10, 33.26, 29.34, 21.48$ e $b_{i2} = -813.04, -53.15, -39.58, -21.35$ respectivamente. Os itens citados acima apresentam parâmetro de discriminação negativo e por isso não apresentam $b_{i1} \leq b_{i2}$, o que nos dá mais indícios de que o modelo possui limitações.

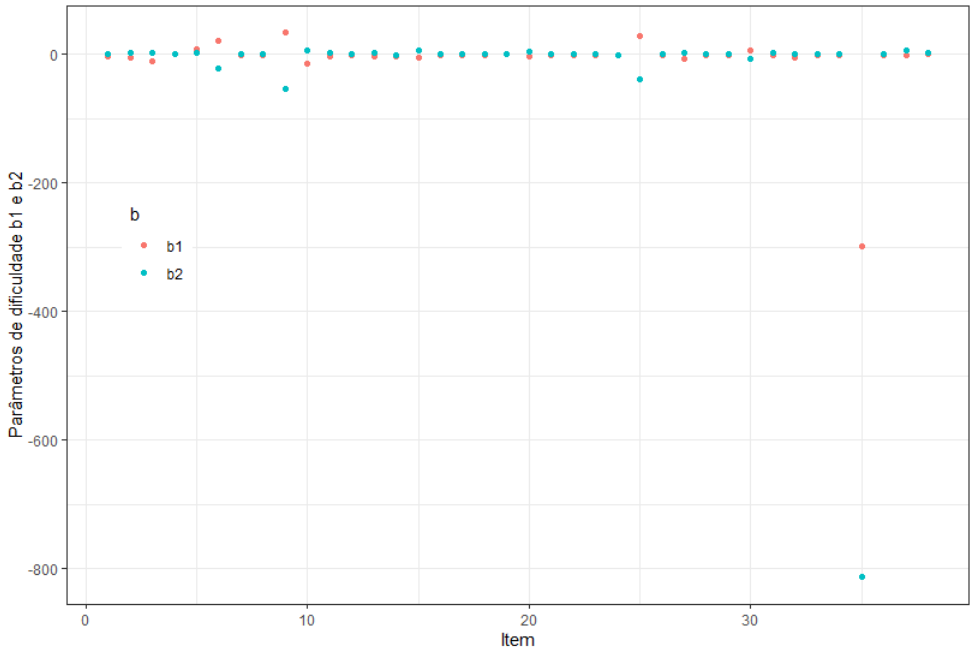

Figura 24: Plot do valor dos parâmetros de dificuldade b_{i1} e b_{i2} de cada item

E finalmente, a figura abaixo nos mostra um gráfico do plot da proficiência estimada via o ML2 versus a proficiência obtida via MRG. Percebesse uma pequena dispersão para proficiências mais baixas e a medida que o valor das proficiências aumenta se cria uma forte semelhança entre os dois modelos. Esta semelhança também pode ser vista pelo coeficiente de correlação de Pearson = 0.79.

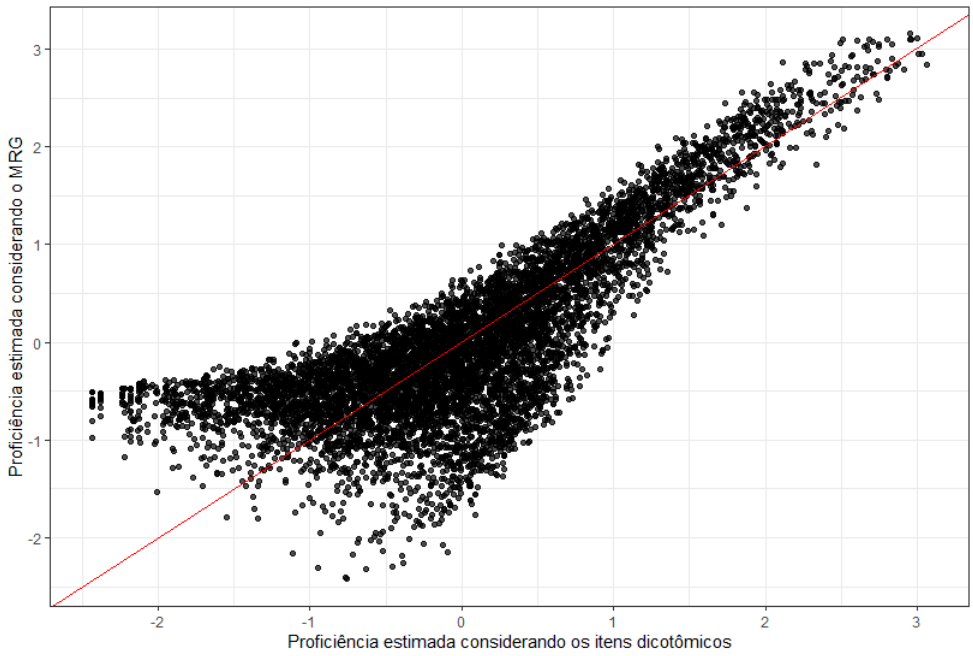

Figura 25: Gráfico da Proficiência estimada considerando o Modelo Dicotomizado versus proficiência estimada considerando o Modelo de Resposta Gradual 1

5.5 Modelo de Resposta Gradual 2

Os diversos valores absurdos e inconsistentes dos parâmetros encontrados no MRG1 indicam que o modelo não está bem ajustado, o que possivelmente foi causado por uma ordenação das categorias de respostas errada. Dito isto, nesta seção iremos propor o MRG2 com uma ordenação de respostas diferente da utilizado no MRG1.

Neste modelo iremos considerar a categoria "não resposta"como a primeira categoria, ou seja, iremos considerar 1=não resposta, 2=errado e 3=certo. A tabela 7 abaixo contém as medidas resumo dos parâmetros e da proficiência estimada no modelo.

Fazendo agora uma breve análise da tabela notamos que o parâmetro de discriminação a_i apresenta valores positivos para todos os itens variando entre 0.47 e 1.82. Já para os parâmetros de dificuldade observamos que os valores são coerentes em b_{i1} variando entre -4.13 e 1.03 e também em b_{i2} variando entre -1.62 e 2.30, e finalmente para as proficiências estimadas θ_j também não observamos nenhum valor absurdo ou discrepante.

Medidas	Parâmetro de Discriminação (a_i)	Parâmetro de Dificuldade (b_{i1})	Parâmetro de Dificuldade (b_{i2})	Proficiência Estimada (θ_j)
Mínimo	0.4755	-4.1347	-1.6195	-3.4186
1 Quartil	0.9619	-1.4868	0.1363	-0.5617
Mediana	1.1828	-0.5825	0.5299	0.0924
Média	1.1712	-0.8512	0.5874	-0.0005
3 Quartil	1.3981	-0.1520	1.0544	0.6921
Máximo	1.8192	1.0302	2.2980	2.7409

Tabela 7: Tabela com medidas padrões dos parâmetros e proficiência do MRG2

Na Figura 26 temos o gráfico da CCIs de todos os itens sendo a linha vermelha desenvolvida com o parâmetro de dificuldade b_{i1} e a linha azul elaborada com b_{i2}. Em ambas as imagens observamos curvas muito mais coerentes do que as obtidas no MRG1. Notamos apenas uma pequena discrepância inicial dos itens 14 e 5 devido ao seus baixos valores de b_{i1} e b_{i2}.

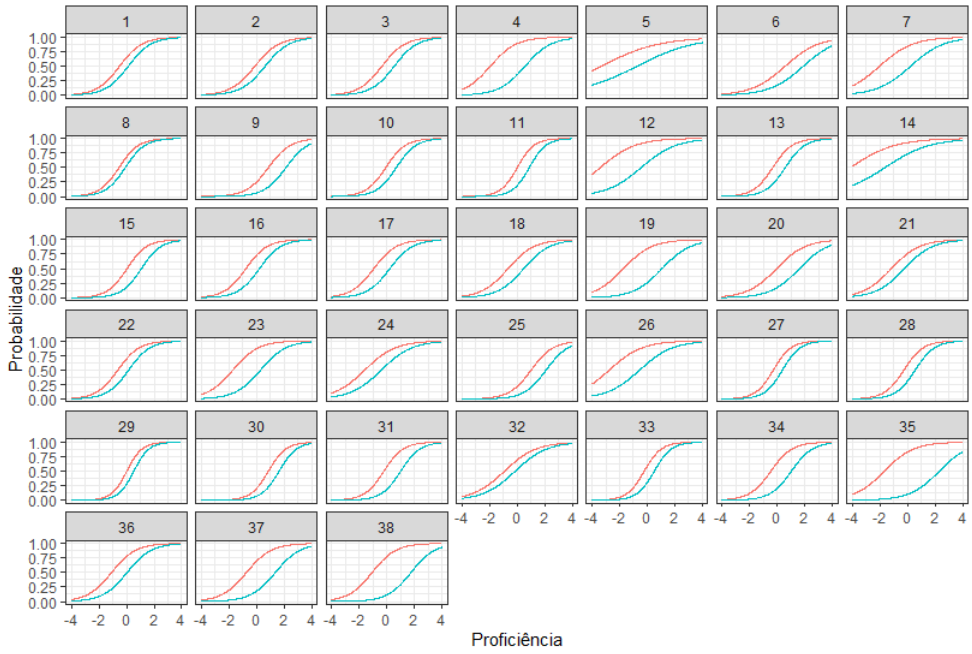

Figura 26: Plot do valor do parâmetro de discriminação a_i de cada um dos itens.

A figura 27 nos mostra o gráfico do parâmetro de discriminação de cada um dos itens. Aqui visualizamos com clareza como todos os itens possuem valor do parâmetro positivo sendo 1.17 o valor médio dos parâmetros.

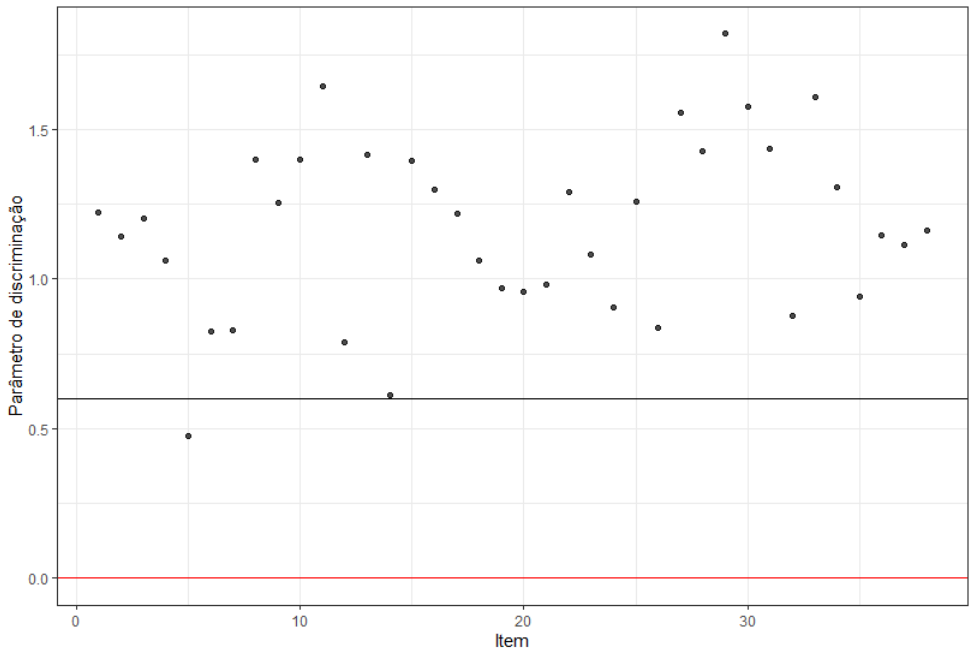

Figura 27: Plot do valor do parâmetro de discriminação a_i de cada um dos itens.

Na figura 28 observamos o gráfico dos parâmetros de dificuldade b_{i1} e b_{i2} de cada item. Os valores dos parâmetros obtidos se apresentam coerentes diferentemente do que foi obtido no MRG1. Também é fácil notar que b_{i2} é maior que b_{i1} para todos os itens.

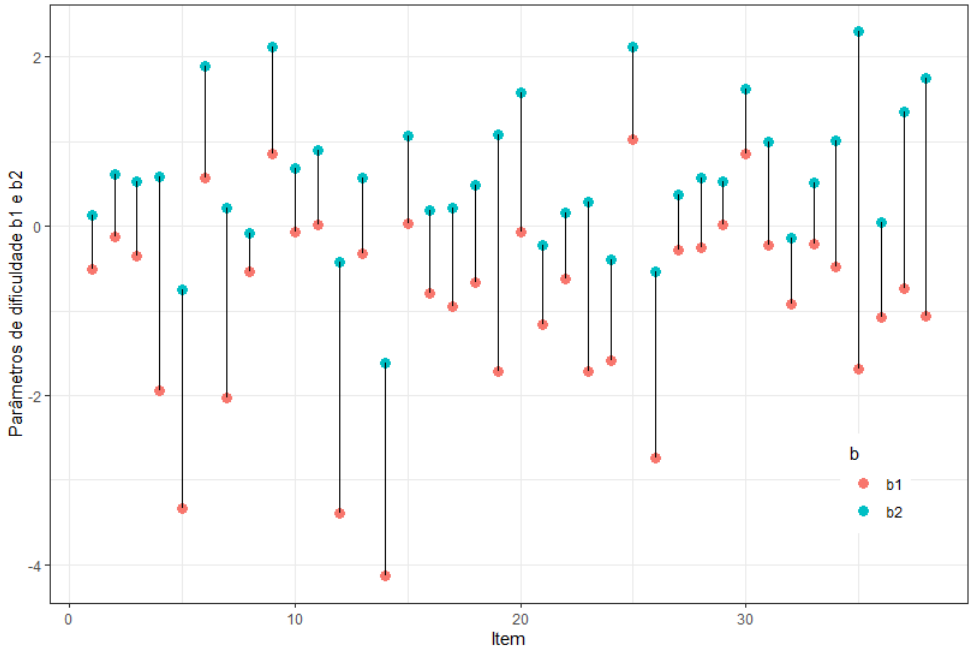

Figura 28: Plot do valor dos parâmetros de dificuldade b_{i1} e b_{i2} de cada um dos itens.

Dadas as análises acima, fica evidente o notável melhor desempenho do MRG2 sobre o MRG1. Para saber o quão melhor é o desempenho do MRG2 os gráficos abaixo visam fazer uma comparação entre os dois modelos MRG.

As figuras 29 e 30 abaixo comparam a proficiência estimada via o MRG1 e o MRG2 com a proficiência estimada via ML2. Entre o MRG1 e o ML2, percebe-se uma grande dispersão para proficiências mais baixas e a medida que o valor das proficiências aumenta se cria uma forte semelhança entre os dois modelos. Porém entre o MRG2 e o ML2 temos uma acentuada correlação ao longo de todos os valores. Estas correlações ficam melhor visualizadas quando utilizamos o coeficiente de correlação de Pearson igual a 0.79 entre o MRG1 e o ML2, e 0.93 entre o MRG2 e o ML2.

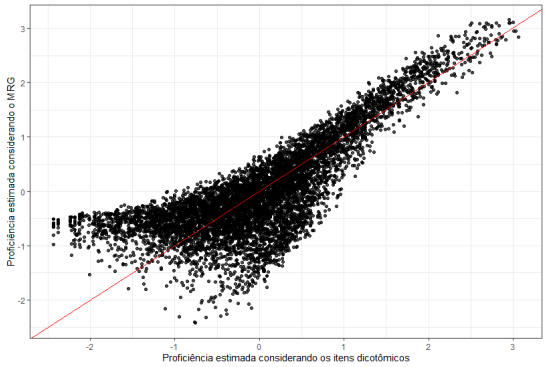

Figura 29: Plot da Proficiência estimada considerando o Modelo Dicotomizado versus a proficiência estimada considerando o Modelo de Resposta Gradual 1

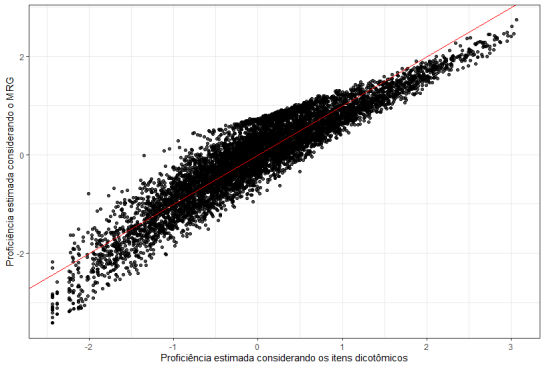

Figura 30: Plot da Proficiência estimada considerando o Modelo Dicotomizado versus a proficiência estimada considerando o Modelo de Resposta Gradual 2

Já as figuras 31 e 32 nos mostram os gráficos de comparação da proficiência estimada via MRG1 e MRG2 com o escore padronizado do vestibular. Notamos que agora o MRG1 apresenta uma correlação significativamente maior com o escore padronizado do que o MRG2. Com os coeficientes de correlação de Pearson iguais a 0.95 e 0.54 respectivamente. Isto ocorre porque o MRG1 adota ordenação para as categorias similar à utilizada no cálculo do escore do vestibular. Para indivíduos com maior número de itens com resposta correta, a correlação entre as estimativas da proficiência com o escore do vestibular é alta para os dois modelos.

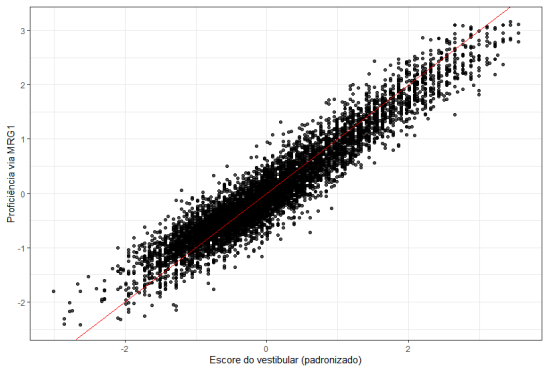

Figura 31: Plot da Proficiência estimada via Modelo de Resposta Gradual 1 versus o escore do vestibular padronizado.

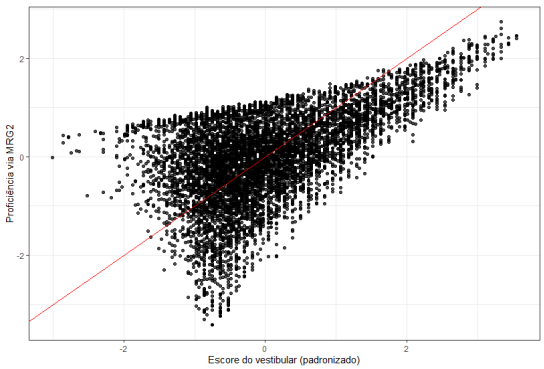

Figura 32: Plot da Proficiência estimada via Modelo de Resposta Gradual 2 versus o escore do vestibular padronizado.

6 Conclusão

O estudo aqui elaborado teve como objetivo analisar as notas das 38 questões do tipo A e C de química do vestibular da UnB de 2014, elaborado pelo CEBRASPE. O trabalho consistiu em comparar o escore convencional utilizado no vestibular com os resultados obtidos através de 4 modelos de TRI: os modelos dicotomizados de dois e três parâmetros e dois modelos de resposta gradual com ordenação de categorias diferentes.

Analisando os itens do vestibular ficou evidente a preferência pela não resposta para itens do tipo A bem como a grande quantidade de resposta erradas em itens do tipo C, provavelmente causada por tentativas de acerto ao acaso por parte dos candidatos.

Ao calcular as proficiências estimadas dos alunos através dos modelos dicotômicos de dois e três parâmetros, foi observado que alunos que não respondem as questões possuem proficiência menor que aqueles que erram, possivelmente porque os candidatos que erram possuem algum conhecimento sobre a questão diferentemente daqueles que não respondem. O modelo de resposta gradual 1 com a ordenação das categorias 1=errado, 2=não resposta e 3=certo apresentou resultados absurdos e inconsistentes, indicando que o modelo não foi corretamente ajustado.

Isso nos levou a elaborar o modelo de resposta gradual 2, com as categorias 1=não resposta 2=errado e 3=certo, modelo que apresentou os melhores e mais coerentes resultados quando comparado com o método convencional, ainda com a vantagem de considerar graus de discriminação e dificuldade para cada um dos itens.

O melhor desempenho do MRG2 é coerente com o observado nas figuras 18 a 21, que indicam que candidatos com maior conhecimento tendem a responder o item e eventualmente erram a resposta, enquanto indivíduos com menor conhecimento tendem a não responder o item, especialmente itens do tipo A.

E finalmente. Os itens 30, 5, 6, 25, 9 e 35 apresentam baixo valor do parâmetro de discriminação no ML3 e ML2. Isto é uma indicação de que estes itens não devem ser utilizados em futuras provas do vestibular da UnB.

Referências

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*, 2000.

ANJOS, A. dos; ANDRADE, D. F. de. Teoria da resposta ao item com uso do r.

BAKER, F. B. *The basics of item response theory*. [S.l.]: ERIC, 2001.

DEMARS, C. *Item response theory*. [S.l.]: Oxford University Press, 2010.

FERNÁNDEZ, J. M. et al. *Teoría de respuesta a los ítems: un nuevo enfoque en la evolución [ie evaluación] psicológica y educativa*. [S.l.]: Pirámide, 1990.

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