Um Método Baseado na Teoria da Resposta ao Item para Avaliação e *Feedback* Automático no Contexto do ENEM

Edwin Monteiro[1], **Raimundo Barreto**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
CEP 69067-005 – Manaus – AM – Brasil

{edwin,rbarreto}@icomp.ufam.edu.br

Abstract. *The assessment of knowledge is a common task in the field of education, whether in assessing the learning or selecting candidates in college admission exams. Exams based on multiple-choice questions such as the Exame Nacional do Ensino Médio (ENEM) do not offer the student significant contributions to the understanding of their performance. This study provides formative feedback to students and teachers through the Item Response Theory (IRT) technique. In this paper we adopted the ENEM database to provide a formative feedback providing guidelines to investigate student's difficulties with the aim to improve the teaching-learning process.*

Resumo. *A avaliação dos conhecimentos é uma tarefa corriqueira no âmbito da educação, seja para avaliar o aprendizado ou selecionar candidatos em vestibulares. Os exames baseados em questões de múltipla escolha como o Exame Nacional do Ensino Médio (ENEM) não oferecem ao estudante contribuições significativas para o entendimento de seu desempenho. Este estudo fornece, tanto para estudantes quanto para professores, um feedback formativo por meio da técnica de Teoria da Resposta ao Item (TRI). Neste artigo adotamos o banco de dados do ENEM para fornecer um feedback formativo fornecendo diretrizes para investigar as dificuldades dos alunos com o objetivo de melhorar o processo de ensino-aprendizagem.*

1. Introdução

A sociedade do século XXI está imersa nos benefícios provindos da tecnologia de modo que esta não depende dos meios tradicionais para obter conhecimento. A geração que nasceu após o advento da internet constrói o conhecimento a partir de objetos de seu interesse, o que dificulta o papel da escola enquanto instituição formadora, pois os alunos desta época não possuem o perfil para o qual o sistema educacional foi originalmente concebido [Giraffa 2013]. Além disso, a abordagem pedagógica tradicional de avaliação do discente, segundo [Caldas e Favero 2009], impõe sobrecarga ao professor durante a correção e dificulta o acompanhamento do processo de aprendizagem do estudante.

Dentre as tentativas de acompanhar os estudantes propõem-se o uso de computadores na educação, uma vez que estes apresentam recursos para auxiliar o processo de mudança da escola, principalmente porque possibilita a criação de ambientes de ensino que enfatizem a construção do conhecimento e não somente a instrução [Valente 2010]. Conforme [Rocha 2007], o computador deve ser utilizado como um meio e não um fim, devendo ser manuseado de maneira a considerar o desenvolvimento dos componentes

DOI: 10.5753/sbie.2022.225182

curriculares. Contudo, apenas implantá-lo na educação não implica em uma melhora satisfatória do ensino. Segundo [Leitão 2017], além dos recursos tecnológicos é preciso estudar o engajamento do aluno no processo de ensino-aprendizagem, buscando meios para avaliar e compreender o nível de entendimento e dificuldades dos estudantes.

A TRI surgiu como alternativa à Teoria Clássica das Medidas, oriunda do trabalho de [Spearman 1961], e que recebeu a contribuição de diversos pesquisadores até a ascensão da TRI, conforme apresentado em [Fletcher 2010]. Dentre as vantagens da TRI estão: (i) facilita a produção, aplicação e correção de exames; (ii) compara traços latentes (processos hipotéticos) de indivíduos em populações diferentes quando submetidos ao mesmo teste que tenha itens comuns; (iii) compara indivíduos na mesma população submetidos a testes distintos [Andrade et al. 2000]; e (iv) proporciona uma pontuação mais justa devido à detecção de questões assinaladas corretamente de modo artificial (adivinhação).

No Brasil, a avaliação automática, nos moldes citados, é aplicada no Exame Nacional do Ensino Médio (ENEM), uma prova que permite o ingresso de pessoas às universidades de acordo com a pontuação obtida. O sistema de correção do ENEM utiliza a Teoria da Resposta ao Item (TRI) proposta por [Lord 1952], a qual prioriza uma avaliação do desempenho de examinandos mediante a identificação de suas habilidades. Assim, o modelo fornece uma probabilidade de acerto por item, dada a habilidade do candidato. Ou seja, por esse modelo, a avaliação automática não está limitada apenas em verificar o assinalamento das respostas, mas, também, em determinar as habilidades que dificilmente são exploradas apropriadamente na correção tradicional.

A principal contribuição deste trabalho está no fornecimento de *feedback* formativo e direcionado aos alunos e professores, indicando aos discentes suas dificuldades em cada tópico, e fornecendo aos docentes informações que ajudem na melhoria do processo de ensino-aprendizagem, a partir da estimação das habilidades dos estudantes, por um modelo logístico, mediante a aplicação de testes objetivos. O *feedback*, com informações pertinentes, ocorre em páginas web geradas automaticamente para cada aluno, o que permite facilidade de acesso e a interação com os artefatos que o compõem.

2. Referencial Teórico

2.1. Avaliação Automática e *Feedback*

A amenização de problemas na avaliação, por meio da avaliação automática, pode ser conduzida de duas formas: avaliar para ajudar a aprender, e avaliar para sintetizar a aprendizagem. Segundo [Santos 2016], a primeira seria o propósito formativo no qual o objetivo é fornecer evidências fundamentadas e sustentadas de forma a agir para apoiar o aluno enquanto a segunda é descrever e dar conta do que o aluno aprendeu. [Van der Kleij et al. 2015] investigaram em uma meta-análise os efeitos de métodos para fornecer *feedback* baseado em questões no contexto da aprendizagem eletrônica. Há três tipos principais de tipos de *feedback* [Pieretti 2015, Van der Kleij et al. 2015]: (i) conhecimento dos resultados (CR), que indica apenas se a resposta assinalada está correta ou incorreta; (ii) conhecimento da resposta correta (CRC), similar ao CR, contudo informa qual a resposta é a correta; e (iii) *feedback* elaborado (FE) que, diferentemente dos anteriores, não há uma distinção clara entre *feedback* em termos de correção e a instrução/sugestão. No FE o processo de fornecer informação agrega tanto características do CR e CRC quanto instruções nas formas de dicas, informações adicionais ou tópicos para

estudo. Este trabalho lida como a geração de *feedback* do tipo elaborado (FE) pois, além de informar o status do assinalamento, sugere o tópico que o aluno deve revisar, assim como outras informações detalhadas nas próximas seções.

2.2. Teoria da Resposta ao Item

A Teoria da Resposta ao Item (TRI) é um modelo matemático em que itens (nomeação dada as questões) são elementos centrais e, portanto, as conclusões não dependem do questionário como um todo e sim de cada item particular que o compõe. A TRI, a partir de seus modelos logísticos, descreve a probabilidade de um estudante assinalar um item corretamente em função dos seus traços latentes, isto é, suas características ou habilidades que não podem ser identificadas diretamente pelo modelo clássico. Esta pesquisa baseia-se no modelo logístico **Rasch** que adota o parâmetro de dificuldade do item. Esse modelo é o mais adequado quando a TRI lida com base de dados com poucos registros, pois a identificação das habilidades permanece concisa. A definição fundamental do modelo logístico foi definida em [Birnbaum 1968]. A Equação 1 descreve a definição simplificada referente ao modelo Rasch: $P(U_{ij} = 1|\theta_j) = \frac{1}{1+e^{-D(\theta_j - b_i)}}$ (1)
com i = 1,2...,I e j = 1,2...,J onde: U_{ij} é uma variável dicotômica que assume o valor 1 quando o indivíduo j responde corretamente ao item i, ou 0 quando o indivíduo j não responde corretamente ao item i; θ_j representa a habilidade (traço latente) do j-ésimo indivíduo; b_i é o parâmetro de dificuldade (ou de posição) do item i, medido na mesma métrica da escala de habilidade θ; 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; e D é um fator de escala, constante e igual a 1. O termo $P(U_{ij} = 1|\theta_j)$, segundo [Baker e Kim 2017], é interpretado como a proporção de respostas corretas para o item i dentre todos os indivíduos de uma população com habilidade θ_j.

Além da descrição numérica, os modelos de TRI podem ser interpretados de maneira visual pelo gráfico da Curva Característica do Item (CCI). Segundo [Baker e Kim 2017], o CCI é definido como um gráfico na forma de uma curva em "S" que descreve a relação entre a probabilidade correta para um item, dado um valor na escala de habilidade. A Figura 1 ilustra a relação entre a probabilidade $P(U_{ij} = 1|\theta_j)$ e os parâmetros do modelo Rasch. O valor b representa um ponto na escala de habilidade cuja probabilidade de resposta correta para um item i é de 50%. Deste modo, quanto maior o valor de b, mais difícil é o item, e vice-versa. O parâmetro a corresponde proporcionalmente ao ponto onde a reta tangente toca a curva. Para o modelo Rash, a será a constante 1. Casos em que a curva se apresenta bastante íngreme, o item tratado é considerado difícil, e portanto, é esperado que apenas indivíduos com habilidades elevadas consigam responder corretamente.

3. Trabalhos Correlatos

Segundo [Chen e Duh 2008], a maioria dos sistemas de educação se concentra no uso de comportamentos, interesses e hábitos do aluno para fornecer serviços personalizados de *e-learning*. Esses sistemas geralmente não consideram a compatibilidade entre a capacidade do aluno e o nível de dificuldade dos cursos recomendados. Visando estimar a capacidade do aluno para serviços de aprendizagem personalizados de acordo com as respostas não nítidas do aluno (ou seja, respostas incertas/imprecisas), [Chen e Duh 2008] apresentam um sistema de tutoria inteligente personalizado com base na Teoria da Resposta a Itens

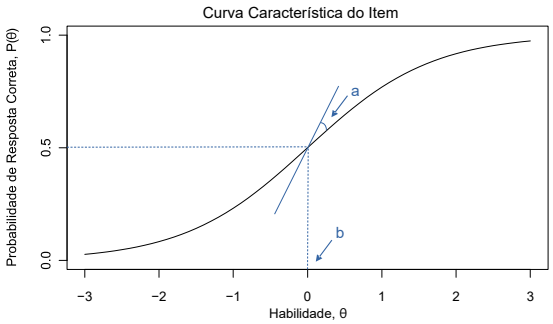


Figura 1. Curva Característica do Item construída a partir dos parâmetros *a* e *b*.

Fuzzy (FIRT), capaz de recomendar cursos com níveis de dificuldade adequados para os alunos, de acordo com as respostas de *feedback* incerto ou *fuzzy* do aluno baseadas em respostas dos alunos às perguntas: como *"Você entende o conteúdo do material do curso?"* e *"Como você pensa sobre a dificuldade dos materiais do curso?"*. Se um aluno puder compreender completamente o material recomendado, o grau de compreensão inferido do aluno estará próximo de um.

Por sua vez, o trabalho de [Yarandi et al. 2013] propõe um sistema de tomada de decisão que usa a aprendizagem adaptativa para identificar os estilos de aprendizagem, habilidades e conhecimento prévio dos estudantes, com a finalidade de determinar de maneira dinâmica conteúdos educacionais que se adéquem as necessidades de cada aluno. Os estilos de aprendizagem, são obtidos pelo uso das ontologias dada a flexibilidade e extensibilidade na modelagem de conceitos e relacionamentos, conforme explicam [Esichaikul et al. 2011]. Os dados coletados durante a interação do estudante com a ferramenta são fornecidos ao modelo da Teoria da Resposta ao Item a fim de determinar as habilidades dos alunos. O sistema permite a aprendizagem adaptativa em dois aspectos: (i) permite a apresentação de conteúdos para diferentes níveis de estudantes com características distintas; e (ii) sugere caminhos de aprendizagem adaptativos, por exemplo, aprender um novo tópico, repetir um tópico com mais detalhes ou fazer mais exercícios com níveis de dificuldade inferiores ou superiores.

Com base nos trabalhos relatados, a principal contribuição desta pesquisa é fornecer um método de avaliação automática que construa *feedback* formativo direcionado aos alunos e professores, com a finalidade de permitir aos alunos compreender de forma clara o porquê de suas dificuldades, tomar conhecimento sobre suas habilidades e expor a relação entre os tópicos compatíveis com o seu grau de habilidade e quais habilidades podem influenciar em uma melhora de seu desempenho. Quanto aos professores, é possível fornecer uma visão geral do desempenho de uma turma destacando as questões mais fáceis, as mais difíceis, a visualização do desempenho individual de cada estudante e a qualidade dos itens elaborados para um exame elaborado, permitindo ao professor uma intervenção mais clara e concisa nas dificuldades de estudantes a fim de melhorar o processo de ensino-aprendizagem.

4. Metodologia

Devido ao grande volume da base de dados do ENEM de 2020, os dados analisados consideram o Índice de Desenvolvimento da Educação Básica (IDEB) de 2021 para o ensino

médio. Assim, optou-se pelo Amazonas por ser o estado brasileiro situado na última colocação conforme o ranking elaborado por [IDEB 2021]. O objetivo do experimento é identificar as habilidades dos estudantes e as dificuldades dos itens para gerar *feedback* formativo que possa ser direcionado tanto aos alunos quanto aos professores.

4.1. Característica dos Dados e Pré-processamento

A base de dados contém registros das quatro áreas do conhecimento: Linguagens, Códigos e suas Tecnologias (LC); Ciências Humanas e suas Tecnologias (CH); Ciências da Natureza e suas Tecnologias (CN); e Matemática e suas tecnologias (MT). Cada área contém 45 itens, totalizando 180 itens objetivos conforme exemplifica a Tabela 1. Além disto, o exame adota um padrão de cores para diferenciar a ordem das questões em cada caderno de prova, portanto levou-se em consideração a prova de cor azul totalizando 53.708 alunos analisados. Alinhado a essas 4 áreas, o [INEP 2022] disponibiliza a matriz de referência do ENEM, um documento que estabelece as competências investigadas em cada área do conhecimento e as habilidades esperadas para cada competência. Diante disso, tabelas auxiliares foram elaboradas para estruturar a relação entre esses termos a fim de auxiliar o processo de correção e posterior elaboração de *feedback*.

Tabela 1. Alguns dos dados contidos da base de dados do ENEM 2019.

ID Aluno	Respostas para LC	Nota em LC
190001028033	BEADE99999BCBDEDDCEEAAACCACEBCCCADCBEBDACDABEC...	548,6
190001028039	ABEBD99999DCCDAEBACDAACDCCBAAAEEDAAAABCEADCBBC...	447,8
190001028106	99999DEEBACCBBCAAADCAACECBECCCACCAAACBCCBDABAA...	473,4
190006119525	99999CCEBEDCEABBAEEEDACECBDEACCACDEEEBACDDEBEC...	500,2
190006119530	BDABE99999CCBABCDEDEDAAECBDCCEDEACCCEBAACDBCBE...	571,0
190006119536	BDABE99999BCBABCDBDCDAAEDAAEBBECBCCDEEAAEDDDBE...	636,3

A correção das provas de cada aluno, conforme dados da Tabela 1, é realizada por meio da construção de uma matriz binária (0,1) por área indicando respectivamente o erro ou o acerto dos estudantes para cada questão a fim de viabilizar o início da etapa de experimentação. A Tabela 2 apresenta a correção binária (erros ou acertos) para uma amostra de 6 alunos da área de Linguagens, Códigos e suas Tecnologias. A tabela ainda contém a coluna "Pontuação Total" que corresponde à soma total dos acertos de cada aluno. Para esta exemplificação a pontuação considera apenas os itens dessa tabela.

Tabela 2. Amostra de respostas dos alunos da área de Linguagens e Códigos.

ID Aluno	Item.1	Item.2	Item.3	Item.43	Item.44	Item.45	Pontuação Total
190001028033	1	0	1	1	0	0	3
190001028039	0	0	0	0	0	0	0
190001028106	0	0	0	0	0	0	0
190006119525	0	0	0	0	0	0	0
190006119530	1	1	1	0	0	0	3
190006119536	1	1	1	1	0	0	4

4.2. Experimentação Utilizando a Base de Dados

Esta seção descreve a etapa de experimentação que visa determinar os parâmetros de dificuldade dos itens e as habilidades dos alunos em uma mesma métrica de habilidade.

Como o modelo TRI não conhece, *a priori*, o traço latente que cada um dos de examinandos possui. A estimação dos parâmetros é realizada em duas etapas chamadas de teste de calibração. Inicialmente o modelo estima a dificuldade dos itens e, em um segundo momento, a partir do parâmetro b, os valores θ das habilidades são estimados. Para uso deste modelo, de acordo com [Baker e Kim 2017], as únicas informações necessárias, sobre a base de dados em análise, são as frequências da pontuação total f e a soma dos itens em cada coluna s. Alguns vetores precisam ser inicializados *a priori*, como é o caso de a, b e θ, que representam o parâmetro de discriminação, o parâmetro de dificuldade do item e o parâmetro de habilidade, respectivamente. Por definição do modelo Rasch, o parâmetro de discriminação a é assinalado com valor 1 para todos os J itens. Já o parâmetro b, segundo [Baker e Kim 2017], é inicialmente definido pela Eq. 2.

$$b = \left(\log\left(\frac{\sum_{n=1}^{|f|}(f_n) - s_1}{s_1} \right), ..., \log\left(\frac{\sum_{n=1}^{|f|}(f_n) - s_j}{s_j} \right) \right), \tag{2}$$

onde $|f|$ corresponde ao total de elementos em f e j que coincide com o valor de J. O parâmetro de habilidade θ é definido pela técnica de ancoragem (*anchoring*) que assinala os mesmos valores do coeficiente de discriminação a para o vetor θ. No segundo momento, θ é estimado pela Eq. 3 onde N é o total de elementos em θ.

$$\theta = \left(\log\left(\frac{1}{J-1} \right), ..., \log\left(\frac{N}{J-N} \right) \right) \tag{3}$$

5. Resultados e Discussões

Na discussão dos resultados obtidos para cada área do conhecimento, as áreas estão organizadas por dia de aplicação da prova, isto é, (i) Linguagens, Códigos e suas Tecnologias e Ciências Humanas e suas Tecnologias; e (ii) Ciências da Natureza e suas Tecnologias e Matemática e suas Tecnologias.

Tabela 3. Amostra dos parâmetros de dificuldade estimados para a prova azul.

Item LC	Dificuldade LC	Item CH	Dificuldade CH	Item CN	Dificuldade CN	Item MT	Dificuldade MT
34	0.78	74	0.77	95	0.49	138	0.57
42	0.74	61	0.61	110	0.44	146	0.46
39	0.66	86	0.59	106	0.31	173	0.30
32	0.03	80	-0.01	126	-0.02	161	0.13
40	-0.02	79	-0.01	108	-0.02	175	0.12
25	-0.49	66	-0.53	120	-0.30	140	-0.60
35	-0.59	59	-0.63	114	-0.43	144	-0.60
19	-0.87	52	-0.64	118	-0.44	136	-0.71

Entre os artefatos gerados estão as dificuldades dos itens (Tabela 3) e as habilidades estimadas (Tabela 4). Além desses, as curvas características, semelhantes a Figura 1, são comentadas ao inter-relacionar os artefatos. Vale destacar que, dada a quantidade de itens por área, a análise dos artefatos leva em consideração uma amostra contendo três valores elevados, dois medianos e dois baixos para fins de comentário.

5.1. Primeiro Dia de Avaliação

De acordo com a [INEP 2022], o item 34 visa *"avaliar a habilidade do aluno em compreender e usar a língua portuguesa como língua materna, geradora de significação e*

Tabela 4. Amostra das habilidades estimadas para as 4 áreas do conhecimento.

Pontuação LC	Habilidade LC	Pontuação CH	Habilidade CH	Pontuação CN	Habilidade CN	Pontuação MT	Habilidade MT
41	2,39	43	3,12	40	2,10	42	2,67
40	2,14	41	2,37	37	1,55	41	2,36
39	1,93	40	2,12	36	1,40	40	2,11
31	0,82	31	0,86	26	0,32	22	-0,04
30	0,72	27	0,04	17	-0,50	21	-0,13
3	-2,70	3	-2,69	3	-2,66	3	-2,68
2	-3,13	2	-3,12	2	-3,09	2	-3,11
1	-3,85	1	-3,84	1	-3,80	1	-3,83

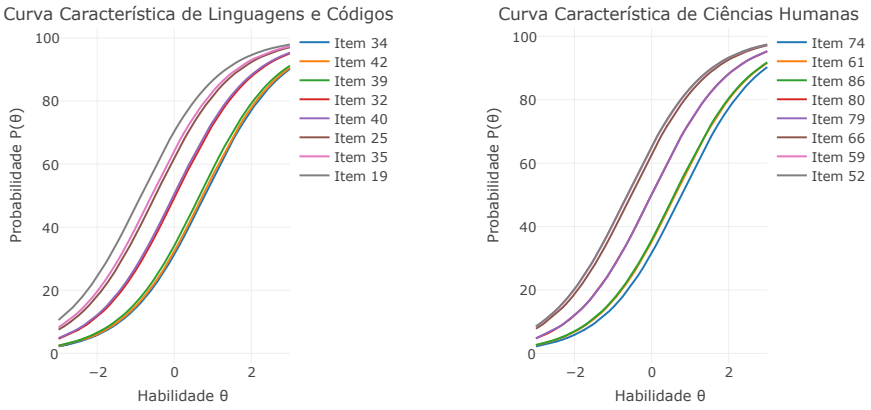
integradora da organização do mundo e da própria identidade". Ao analisar o item da Tabela 3, é possível notar que esta é a questão mais difícil de LC. A sua dificuldade está estimada em $0,78$. Ao correlacionar esse valor com a quantidade de acertos é possível compreender que a estimativa está de acordo, pois dos 53.708 alunos apenas 12.328 acertaram o item, isto é, 23% de acerto. Avaliando a curva de cor azul na Figura 2a, percebe-se que é a curva de maior habilidade, pois conforme a Seção 2, os valores de habilidade e dificuldade residem na mesma escala métrica. Assim, quanto maior a dificuldade, mais à direita está a curva e, portanto, maior é a habilidade requerida. Para este caso, é necessária uma habilidade de pelo menos $0,78$ para o aluno ter a probabilidade de 50% de acerto. No entanto, entre as habilidades identificadas na Tabela 4, apenas os alunos com habilidade entre $0,82$ e $2,39$ estariam aptos para responder à questão corretamente.

A Equação 4 do modelo Rasch ajuda a entender as chances de acerto.

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-1(0,82-0,78)}} = \frac{1}{1 + e^{-0,037}} = 0,5099 \tag{4}$$

Atribuindo $0,82$ e $0,78$ para θ_j e b_i, a probabilidade de assinalamento correto é de 51%. Se a habilidade θ for substituída pela estimativa inferior, $0,72$, então a probabilidade é de apenas $48,50\%$. Logo, é possível afirmar que apenas alunos com habilidade de 0,82 possuem probabilidade concreta de acertar o item. Por outro lado, analisando o item 19, que apresenta a maior discrepância em relação aos demais itens, pois dos 53.708, apenas 31.013 (57,74%) alunos assinalaram o item corretamente. Em termos de CCI (curva em cinza mais à esquerda), é possível validar pela Figura 2a que os alunos tiveram facilidade no item. De acordo com as informações da matriz de referência, O tópico investiga se *"o aluno é capaz de entender os princípios, a natureza, a função e o impacto das tecnologias da comunicação e da informação em sua vida pessoal e social, no desenvolvimento do conhecimento, associando-o aos conhecimentos científicos"*. O item ser mais fácil é esperado pelo fato que a nova geração estar em constante contato com as variadas ferramentas tecnológicas o que abre margem para um maior conforto na resolução do item. Esta característica impactou diretamente em uma menor estimativa de dificuldade, $-0,87$. De forma geral, as curvas representam os dados expressos na Tabela 3 onde itens 34, 42 e 39 estão bem próximos, os itens medianos mantêm uma proximidade em termos de dificuldade e, por fim, fica nítido que o item 19 é o mais fácil estando em evidência.

No que diz respeito a área de Ciências Humanas e suas Tecnologias, é possível compreender que os níveis de dificuldade estão distribuídos uniformemente, o que facilita o acompanhamento pela Tabela 3. A diferença mais perceptível entre os níveis de dificuldade está nos itens 61, 74 e 86, o conjunto de itens mais difíceis, estando as curvas



(a) CCI para LC. **(b)** CCI para CH.

Figura 2. Curvas características do primeiro dia de prova: LC e CH.

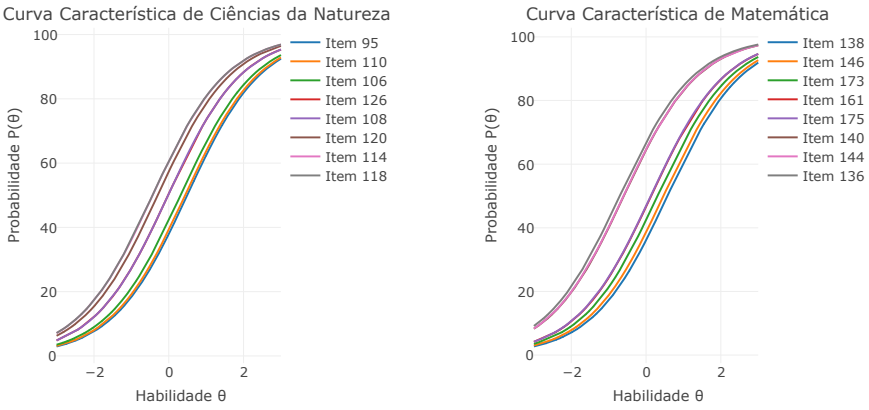
bem próximas. As curvas para os itens 79 e 80, e curvas 52 e 59, apresentam-se como uma única curva no plano, pois têm dificuldade mediana estimadas muito próximas. Já o item 66 possui dificuldade de -0,53. Em linhas gerais, um aluno mediano com a habilidade mais próxima, cerca de 0,04, teria 51% de chances para assinalar corretamente os itens 79 e 80. Outro ponto a ser considerado está na descrição das habilidades onde o item 79 visa *"utilizar os conhecimentos históricos para compreender e valorizar os fundamentos da cidadania e da democracia, favorecendo uma atuação consciente do indivíduo na sociedade"* e o item 80 avalia *"compreender a produção e o papel histórico das instituições sociais, políticas e econômicas, associando-as aos diferentes grupos, conflitos e movimentos sociais"*. Assim, era esperado pelo professor que o aluno pudesse estabelecer uma conexão entre as habilidades de modo a fazer uma marcação adequada dos itens, dados os conhecimentos prévios do estudante. A respeito do item 74, por meio da dificuldade estimada em 0,77, um aluno com habilidade mediana não tem chances claras de acertar (veja a Tabela 3). Assim, o item acaba delimitando o intervalo de habilidades entre [0,86; 3,12] onde o primeiro garante 52% e o segundo 91% de chances para acerto. Vale destacar que a habilidade está diretamente relacionada com o total de acertos, assim o aluno precisaria ter, pelo menos, 31 acertos o que corresponde aos 0,86 de habilidade.

5.2. Segundo Dia de Avaliação

O segundo dia de avaliação conta com curvas bastante distintas por área. Na Figura 3a é possível definir quais os níveis de dificuldade. Ao relacionar os itens com as respectivas dificuldades da Tabela 3, os itens 126 e 108, embora medianos, registram 73% de assinalamentos incorretos, com uma dificuldade de -0,02. Este fato pode indicar ao professor que a questão não se comportou conforme planejado. As três questões mais difíceis representam uma média de assinalamento incorreto em 86%. Ao considerar as dificuldades em ordem decrescente $0,49$; $0,44$ e $0,31$ nota-se que apenas os alunos que assinalaram 26 itens corretos estariam aptos, considerando a habilidade, para responder as questões mais difíceis o que representa apenas 0,8% da população. Esse percentual estabelece um alerta para que os conteúdos sobre *"ciclos biogeoquímicos"*, *"mecanismos de transmissão da vida"* e *"associação entre informações nas diferentes formas de linguagem utilizadas em ciências físicas, químicas ou biológicas"* tenham prioridade nos estudos dos estudantes que tiveram menos de 26 acertos em CN.

Por fim, em relação aos itens de Matemática e suas Tecnologias (Figura 3b) não há uma clara separação entre dificuldade mediana e elevada, o que denota uma maior dificuldade dos estudantes nesta área. A Tabela 3 permite compreender que as curvas dos três itens que se apresentam como fáceis diferem de 0,11 sendo que as os itens 140 e 144 possuem o mesmo parâmetro de dificuldade e avaliam o mesmo tópico: *"Reconhecer, no contexto social, diferentes significados e representações dos números e operações - naturais, inteiros, racionais ou reais"*. Por sua vez, o item 136 avalia a *"interpretação da localização e da movimentação de pessoas/objetos no espaço tridimensional e sua representação no espaço bidimensional"*. Ou seja, os tópicos não demandam tanta atenção dos estudantes, pois uma habilidade igual ou superior a -0,60 poderia assinalar todos os referidos itens corretamente. Para as dificuldades mediana e elevada, o cenário não é trivial porque há muitas habilidades próximas entre si, o que dificulta a identificação de quais itens devem demandar maior atenção do estudante. É possível sugerir um estudo segmentado em partes de acordo a necessidade de conhecimento para cada tópico.

O conteúdo dos itens 146, 173 e 161 estão relacionados com tópicos estatísticos. Consultando em [INEP 2021] e [INEP 2022], o estudo para desenvolver as habilidades deve abordar a *"análise de informações expressas em gráficos ou tabelas como recurso para a construção de argumentos"*, *"cálculo de medidas de tendência central ou de dispersão de um conjunto de dados expressos em uma tabela de frequências ou em gráficos"*. Já os itens 138 e 175 podem ser estudados de maneira isolada por abordarem conceitos distintos para os demais itens como *"propostas de intervenção na realidade utilizando conhecimentos numéricos"* e *"resolução de situação-problema que envolva conhecimentos geométricos de espaço e forma"*. Por fim, a ordem de estudos para os itens mais difíceis deve prevalecer o item 138 por ter o maior percentual de erros 85%, seguido do estudo dos itens 146, 173 e 161 com 82% de erros e, finalmente, o item 175 com 80% de erro.



(a) CCI para CN. **(b)** CCI para MT.

Figura 3. Curvas características para o primeiro dia de prova: CN e MT.

5.3. Análise do *Feedback* Automático

A TRI foi utilizada não apenas com informações sobre o desempenho de estudantes, mas também com subsídios para que o professor analise se os itens elaborados seguem o propósito da avaliação, isto é, se os itens de um exame cumprem o papel de distinguir o nível de habilidade dos discentes e permitir a identificação de eventuais equívocos na metodologia. O professor pode verificar o nível de dificuldade estimado para um item ao analisar

a curva CCI. Se probabilidade de acerto está muito elevada para habilidades medianas (abaixo de zero), então o item foi considerado fácil pelos examinandos. Analisando outro cenário, no qual muitos alunos erraram um item, a curva característica deve refletir com precisão que a questão avaliada é difícil. A depender do contexto o professor pode tomar atitudes cabíveis para contornar a situação com base nos artefatos gerados.

O *feedback* fornecido ao aluno apresenta o estado do assinalamento, a habilidade, a dificuldade do item, a sua probabilidade atual, o que é investigado na questão, além de ser possível simular, com base nas Figuras 2 e 3 qual seria a habilidade ideal para que o aluno obtivesse um melhor desempenho. O *feedback* professor é semelhante ao destinado para o aluno com o acréscimo das tabelas geradas durante o experimento e analisadas na Seção 5, como a Tabela 2 de assinalamentos, a Tabela 3 de dificuldade dos itens e os gráficos, como o da Figura 2, que contém todas as curvas características para uma determinada disciplina. A ideia deste artigo é que o *feedback* automático sumarize os artefatos em páginas web por disciplina, o que favorece o uso por professores de diferentes áreas do conhecimento. Portanto, o método pode ser aplicado no contexto educacional em situações nas quais o professor elabora uma avaliação/prova escolar e opta por um método de correção automática dos itens que forneça *feedback* elaborado.

6. Conclusão

O uso da TRI em questões de múltipla escolha, como técnica para auxiliar a avaliação dos discentes e elaborar os artefatos de *feedback* automático, permite que informações antes não capturadas pela Teoria Clássica das Medidas, como as habilidades de cada indivíduo e as dificuldades das questões, possam ser identificadas para cada estudante. O *feedback* elaborado proporciona uma comunicação direta e objetiva entre alunos e professores por meio de acesso simples, didático às informações obtidas a partir da análise de rendimento dos estudantes. Os artefatos de *feedback* construídos com base nas estatísticas geradas pela TRI, podem permitir ao aluno uma melhor compreensão do porquê do seu desempenho simplesmente analisando a relação entre sua habilidade estimada e a dificuldade da questão. A correlação entre a dificuldade do item e a habilidade avaliada também permite o estudo de conteúdo compatível com sua habilidade. Do ponto de vista do professor, as informações apresentadas otimizam o seu tempo, isto é, há diminuição na sobrecarga imposta ao professor, de modo que o tempo economizado em correções pode ser destinado para a pesquisa ou criação de novos materiais. Do ponto de vista metodológico, o professor pode atestar se os itens aplicados para avaliar o conhecimento dos examinandos estão de acordo com o nível da turma, e aplicar intervenções quando necessárias.

Para trabalhos futuros, há o interesse de acrescentar outros modelos logísticos da TRI, os quais podem fornecer informações como a discriminação de itens e o acerto casual (chute). Estas informações ajudam no incremento da gama de informações destinadas ao professor, porque as curvas características podem destacar se determinado item é bom em discriminar os alunos que têm a possibilidade de acertar ou não, além de ilustrar e penalizar respostas por acerto casual. Um ambiente web do tipo *dashboard* on-line sintetizando todos os artefatos apresentados neste artigo, está em desenvolvimento visando lidar com itens compatíveis com as diretrizes do ENEM. Pretende-se torná-lo público em breve, assim professores podem acessar o ambiente e inserir suas próprias bases de dados para que o modelo construa os artefatos de *feedback* automático visando auxiliar os estudantes facilitando assim o processo de ensino-aprendizagem.

Referências

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

Baker, F. B. e Kim, S.-H. (2017). *The basics of item response theory using R*. Springer.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Caldas, V. M. e Favero, E. L. (2009). Uma ferramenta de avaliação automática para mapas conceituais como auxílio ao ensino em ambientes de educação a distância. *Simpósio Brasileiro de Informática na Educação-SBIE*.

Chen, C.-M. e Duh, L.-J. (2008). Personalized web-based tutoring system based on fuzzy item response theory. *Expert systems with applications*, 34(4):2298–2315.

Esichaikul, V., Lamnoi, S., e Bechter, C. (2011). Student modelling in adaptive e-learning systems. *Knowledge Management & E-Learning: An International Journal*, 3(3):342–355.

Fletcher, P. R. (2010). Da teoria clássica dos testes para os modelos de resposta ao item. *Rio de Janeiro: Escola Nacional de Ciências Estatísticas*.

Giraffa, L. M. M. (2013). Jornada nas escol@s: A nova geração de professores e alunos. *Tecnologias, Sociedade e Conhecimento*, 1(1):100–118.

IDEB (2021). Índice de desenvolvimento da educação básica. `http://ideb.inep.gov.br/resultado/resultado/resultado.seam?cid=475939`. Acessado em: 20-06-2022.

INEP (2021). Microdados do enem 2021. `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem`. Acessado em: 01-06-2022.

INEP (2022). Matriz de referência do enem. `https://download.inep.gov.br/download/enem/matriz_referencia.pdf`. Acessado em: 03-06-2022.

Leitão, G. d. S. (2017). Uma plataforma de suporte ao docente no contexto da educação digital. UFAM, Universidade Federal do Amazonas.

Lord, F. (1952). A theory of test scores. *Psychometric monographs*.

Pieretti, A. A. R. (2015). Efeito da variação do feedback e da possibilidade de repetição de itens incorretos no desempenho em uma instrução programada. Programa de estudos pós-graduados em psicologia experimental: Análise do comportamento, Pontifícia Universidade Católica de São Paulo.

Rocha, F. E. L. d. (2007). *Avaliação da aprendizagem: uma abordagem qualitativa baseada em mapas conceituais, ontologias e algoritmos genéticos*. Centro tecnológico, Universidade Federal do Pará, Brasil.

Santos, L. (2016). A articulação entre a avaliação somativa e a formativa, na prática pedagógica: uma impossibilidade ou um desafio? *Ensaio: avaliação e políticas públicas em Educação*, 24(92):637–669.

Spearman, C. (1961). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Valente, J. A. (2010). O computador auxiliando o processo de mudança na escola. *NIED-UNICAMP e CED-PUCSP*.

Van der Kleij, F., Feskens, R., e Eggen, T. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, pages 475–511.

Yarandi, M., Jahankhani, H., e Tawil, A.-R. (2013). Towards adaptive e-learning using decision support systems. *International Journal of Emerging Technologies in Learning*.