Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 24/03/2021; 1st round notif.: 12/05/2021; New version: 24/06/2021; 2nd round notif.: 01/07/2021;
Camera ready: 16/07/2021; Edition review: 20/07/2021; Available online: 30/07/2021; Published: 30/07/2021;

Um Método Baseado na Teoria da Resposta ao Item para Avaliação e *Feedback* Automático no Contexto de Educação Digital

Title: A Method Based on Item Response Theory for Automatic Assessment and Feedback in the Context of Digital Education

Edwin Juan Lopes Barboza Monteiro
Universidade Federal do Amazonas
edwin@icomp.ufam.edu.br

Gabriel de Souza Leitão
Universidade Federal do Amazonas
gabriel.leitao@icomp.ufam.edu.br

Raimundo da Silva Barreto
Universidade Federal do Amazonas
rbarreto@icomp.ufam.edu.br

Resumo

A avaliação dos conhecimentos de estudantes é tarefa corriqueira no âmbito da educação, seja para avaliar o aprendizado ou mediar a seleção de candidatos em vestibulares. Os exames que empregam uma avaliação baseada em questões de múltipla escolha, onde o aluno assinala itens durante o tempo de prova e, depois, recebe um feedback sobre a sua nota, não oferecem ao estudante contribuições significativas para o entendimento de seu desempenho. O objetivo geral deste estudo é fornecer, tanto para estudantes quanto para professores, um feedback formativo a partir da estimação das habilidades do examinando e dificuldade dos itens, por meio de uma técnica estatística denominada Teoria da Resposta ao Item (TRI). Esta técnica produz um modelo de dados sobre o desempenho dos estudantes baseado na análise das respostas coletadas nos exames. Um experimento foi realizado utilizando dados de um exame aplicado em uma escola pública de ensino médio do Amazonas. Os resultados obtidos indicam que há uma quantidade valiosa de informações que permitem analisar a relação entre os diversos itens e as habilidades estimadas. É possível classificar os alunos em uma escala de habilidade de modo que o próprio discente pode localizar sua posição em uma disciplina e simular quais tópicos podem ser estudados para obter uma maior habilidade. Do ponto de vista do professor é possível analisar quais tópicos foram mais difíceis para a turma, assim como analisar se um item elaborado está em conformidade com as aulas ministradas. Caso contrário, o professor pode refazer o item ou alterar o nível de dificuldade. Os resultados proporcionaram informações significativas que permitem a elaboração um feedback formativo capaz de fornecer aos examinandos e avaliadores as diretrizes necessárias para investigar dificuldades e contribuir para um melhor rendimento no processo de ensino-aprendizagem.

Palavras-chave: *Feedback Formativo; Teoria da Resposta ao Item; Avaliação Automática; Ambiente Virtual de Aprendizagem; TRI; AVA.*

Abstract

The assessment of students' knowledge is a common task in the field of education, either to assess learning or to mediate the selection of candidates in entrance exams. Exams that employ an assessment based on multiple choice questions, in which the student marks items during the exam time and then receives feedback on his grade, do not offer the student significant contributions to understanding his performance. The general objective of this study is to provide, for both students and teachers, formative feedback from the estimation of the ability of students and difficulty of the items, through a statistical technique called Item Response Theory (IRT). This technique produces a model of data on student performance based on the analysis of the responses collected in the exams. An experiment was carried out using data from an exam applied in a public high school in Amazonas. The results obtained indicate that there is a valuable amount of information that allows analyzing the relationship between the various items and

Cite as: Monteiro, E. J. L. B., Leitão, G. d. S., & Barreto, R. d. S. (2021). A Method Based on Item Response Theory for Automatic Assessment and Feedback in the Context of Digital Education (Um Método Baseado na Teoria da Resposta ao Item para Avaliação e Feedback Automático no Contexto de Educação Digital). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação – RBIE), 29, 746-774. DOI: 10.5753/RBIE.2021.29.0.746.

the estimated abilities. It is possible to classify students on a hability scale so that the student himself can locate his position in a discipline and simulate which topics can be studied to obtain greater hability. From the teacher's point of view, it is possible to analyze which topics were more difficult for the class, as well as to analyze whether an elaborated item is in accordance with the classes taught. Otherwise, the teacher can redo the item or change the difficulty level. The results provided significant information that allows the elaboration of formative feedback capable of providing the examinees and evaluators with the necessary guidelines to investigate difficulties and contribute to a better performance in the teaching-learning process.
Keywords: *Formative Feedback; Item Response Theory; Automatic Evaluation; Virtual Learning Environment; IRT; VLE.*

1 Introdução

A sociedade do século XXI está imersa nos benefícios provindos da tecnologia de modo que não depende, por exemplo, dos meios de comunicação 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, de acordo com Giraffa (2013), não possuem o perfil para o qual o sistema educacional foi concebido.

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. Cabe à escola encontrar meios para acompanhar esse grande desafio que, segundo Santo, Castelano, e Almeida (2012), é constituído de um espaço de mediação entre o aluno e o mundo tecnológico. Entre as tentativas de acompanhar os estudantes está o uso de computadores na educação. Para Valente (2010), o uso do computador apresenta recursos importantes para auxiliar o processo de mudança da escola, pois possibilita a criação de ambientes de ensino que enfatizem a construção do conhecimento e não a instrução.

Entretanto, o mau uso do computador como ferramenta intermediadora do processo de ensino inibe sua capacidade de atuar como ferramenta pedagógica. 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 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.

Pesquisas como as de Caldas e Favero (2009), Isotani e de Oliveira Brandão (2004) e Juniwal (2014) tentam reduzir a sobrecarga imposta ao professor usando a técnica de Avaliação Automática. Uma característica dessa avaliação é fornecer *feedback* imediato após a aplicação de determinado exame. Este *feedback* proporciona ao estudante o acompanhamento do seu desempenho nas avaliações, além de refletir sobre o próprio aprendizado (Iahad, Dafoulas, Kalaitzakis, e Macaulay, 2004). 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 TRI surgiu como alternativa à Teoria Clássica das Medidas, uma metodologia do início do século 20 oriunda do trabalho de Spearman (1961), e que recebeu a contribuição de diversos pesquisadores até a ascensão da TRI, conforme apresentado em Fletcher (2010). Entre as vantagens da TRI estão a facilidade de produzir, aplicar e corrigir exames conforme descrito nos itens abaixo:

- Comparações entre traços latentes (processos hipotéticos) de indivíduos em populações diferentes quando submetidos ao mesmo teste que tenha itens comuns;
- Comparação de indivíduos na mesma população submetidos a testes distintos (Andrade, Tavares, e Valle, 2000);
- Pontuação mais justa devido à detecção de questões assinaladas corretamente de modo artificial (adivinhação).

Trabalhos como os de Caldas e Favero (2009), Juniwal (2014) e Moreira e Favero (2009), utilizam a avaliação automática com *feedback* para criar soluções que facilitem o processo de ensino-aprendizagem. Todavia, o *feedback* entregue ao aluno é puramente quantitativo. Há, entretanto, a necessidade de fornecer mais informações que permitam ao discente entender as dificuldades cognitivas, de tal forma que ele possa aperfeiçoar e construir conhecimentos.

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

Esta seção apresenta os conceitos fundamentais abordados neste trabalho a fim de proporcionar um embasamento teórico que permita a compreensão de termos como Ambientes Virtuais de Aprendizagem, Avaliação Automática e técnicas de *Feedback*, Teoria da Resposta ao Item e Teste de Calibração.

2.1 Ambientes Virtuais de Aprendizagem (AVAs)

O termo Ambiente Virtual de Aprendizagem (AVA) refere-se aos sistemas de aprendizagem eletrônica que atuam como o elo para conectar alunos e professores a um ambiente de educação

que permita experiências similares àquelas vivenciadas em salas de aula tradicionais, porém dispondo de um conjunto de recursos tecnológicos para facilitar o processo de ensino-aprendizagem. No meio acadêmico, a avaliação de discentes é o processo que visa atestar o conhecimento do indivíduo sobre determinado conteúdo.

Segundo Meyer e Mont'Alverne (2021), os AVAs fornecem uma solução integrada para gerir o aprendizado on-line, fornecendo mecanismos de avaliação e acesso a recursos disponíveis que permitem uma aprendizagem significativa. Nesses espaços, encontram-se uma série de ferramentas e recursos que possibilitam aos docentes organizar atividades e tarefas. Existem desde opções mais interativas, como fóruns e chats, até outras mais estruturadas e predefinidas como os questionários. É dentro dos AVAs que se concentra toda ou a maior parte do ensino-aprendizagem (Veloso, 2021).

Os AVAs podem ser empregados como ferramentas de suporte para o Ensino à Distância (EaD), bem como servir de apoio às atividades presenciais em sala de aula ou diferentes ambientes por meio da internet ou rede interna de computadores. Dentre as ferramentas mais comuns na literatura que dão suporte ao ensino à distância pode-se citar: Blackboard[1], Moodle[2] e Sakai[3].

2.2 Avaliação Automática e *feedback*

De acordo com Caldas e Favero (2009), o processo de avaliação impõe sobrecargas ao docente, pois turmas com um número elevado de alunos impossibilitam que o professor obtenha uma visão geral do desempenho da classe e ainda dificultam o acompanhamento individual dos estudantes, já que o professor dedica grande parte do seu tempo à correção das avaliações. O problema torna-se maior quando alunos e professores atuam em um ecossistema de interação à distância, isto é, que não proporciona o acompanhamento e a comunicação direta de ambos os lados no processo de ensino-aprendizagem. Henderson, Ajjawi, Boud, e Molloy (2019) destacam que os estudantes relatam muita insatisfação, pois não conseguem o que desejam com os comentários que recebem sobre o desempenho, já os professores ficam preocupados que os alunos não se envolvam e assim os docentes questionam se o esforço de ensino vale a pena.

AVAs como os citados na Subseção 2.1 têm apoiado o ensino na modalidade EAD através de serviços, como a correção automática de questões objetivas, com o intuito de amenizar problemas na avaliação que, segundo Santos (2016), pode ser feito de duas formas: avaliar para ajudar a aprender, e avaliar para sintetizar a aprendizagem. 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 que a segunda é descrever e dar conta do que o aluno aprendeu.

A avaliação automática deve, portanto, fornecer ao aluno e professor, recursos que permitam não apenas obter a pontuação como indicativo de desempenho, mas também subsídios para que o indivíduo compreenda os sucessos e insucessos. Segundo Pimentel, Real, Braga, e Botelho (2020), compreender os insucessos é essencialmente importante, pois retroalimenta o processo de ensino-aprendizagem podendo, se necessário, reorganizar a sequência do curso, os materiais e os processos avaliativos.

[1] https://www.blackboard.com/pt-br
[2] https://moodle.org/?lang=pt_br
[3] https://www.sakailms.org

Van der Kleij, Feskens, e Eggen (2015) investigaram em uma meta-análise os efeitos de métodos para fornecer *feedback* baseado em questões no contexto da aprendizagem eletrônica. Entre os variados tipos de *feedback*, que podem ser consultados em (Pieretti, 2015), Van der Kleij et al. (2015) descrevem três tipos principais: conhecimento dos resultados (CR) que indica apenas se a resposta assinalada está correta ou incorreta; conhecimento da resposta correta (CRC), similar ao CR, contudo informa qual a resposta é a correta; e por fim, *feedback* elaborado (FE) que, diferentemente dos anteriores, não há uma distinção clara entre *feedback* (em termos de correções) 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.

2.3 Teoria da Resposta ao Item

Os resultados obtidos a partir da aplicação de exames para a avaliação e seleção de estudantes, ainda que automáticos, são expressos, na maioria dos casos, por escores brutos ou padronizados em alguma escala. Além disso, a avaliação e interpretação dos resultados é dependente do conjunto de questões ou itens selecionados para o exame. Tal fato implica em análises associadas à prova como um todo, ou seja, a pontuação do candidato depende da quantidade de questões assinaladas corretamente, o que é uma característica da Teoria Clássica das Medidas (TCM). Deste modo, de acordo com Andrade et al. (2000), a comparação entre candidatos de uma população apresenta uma deficiência, pois é válida somente se estes são submetidos ao mesmo exame.

Para suprir as limitações do modelo TCM, como a proposta neste trabalho, além de outras apresentadas no trabalho de Petrassi, Bornia, e Andrade (2021), foi desenvolvida a Teoria da Resposta ao Item (TRI), 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. Dentre as vantagens do TRI, quando comparado com o TCM, pode-se destacar a capacidade de comparar indivíduos de populações distintas a partir de seus traços latentes, ou seja, a comparação é feita por habilidades presentes em uma escala quando estes são submetidos a questionários que compartilhem itens em comum. De maneira análoga, é possível comparar indivíduos da mesma população quando submetidos a questionários com itens distintos. Segundo Araujo, de Andrade, e Bortolotti (2009), as comparações são possíveis porque tanto os itens quanto os traços latentes residem no mesmo espaço métrico, denominado escala de habilidade. Desta forma, a habilidade é algo comum ao indivíduo e ao item que a exige para que seja assinalado corretamente.

2.3.1 Modelos de TRI para itens dicotômicos

Dentre os modelos de TRI para itens dicotomizados, isto é, itens assinalados como corretos ou incorretos, destacam-se os de 1, 2 e 3 parâmetros. Estes modelos levam em consideração as

respectivas características:

(1) Contém apenas o parâmetro de dificuldade do item;

(2) Contém a dificuldade do item e o parâmetro de discriminação do item;

(3) Contém a dificuldade do item, a discriminação do item e o parâmetro de acerto casual.

Esta pesquisa dá ênfase ao 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 Equação 1 descreve o modelo logístico de 3 parâmetros do qual o modelo Rasch é deduzido e descrito nos parágrafos seguintes.

Modelo Logístico de 3 Parâmetros – ML3: A definição do ML3 desenvolvida por Birnbaum (1968) é dada por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}} \quad (1)$$

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. A CCI é descrita com detalhes na Subsubseção 2.3.2.

c_i é o parâmetro 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 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.

Para obter o modelo Rasch a partir do ML3 basta assinalar o valor 1 para o parâmetro de discriminação do item e assinalar o valor 0 ao parâmetro de acerto casual. A Equação 2 descreve o modelo Rasch.

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-D(\theta_j - b_i)}} \quad (2)$$

2.3.2 Interpretação gráfica do modelo

Além da descrição numérica, os modelos de TRI podem ser interpretados de maneira visual pelo gráfico da Curva Característica do Item. Baker e Kim (2017) definem 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. De forma matemática, a probabilidade de resposta correta é próxima de zero quando os valores de habilidade ficam mais próximos de $-\infty$ e aumenta quando os valores de habilidade aproximam-se de $+\infty$. Na prática, os limites inferior e superior variam entre -4 e 4.

A Figura 1 ilustra a relação entre a probabilidade $P(U_{ij} = 1|\theta_j)$ e os parâmetros do modelo ML3. O parâmetro *b* denota um ponto em alguma região desta escala e representa a habilidade

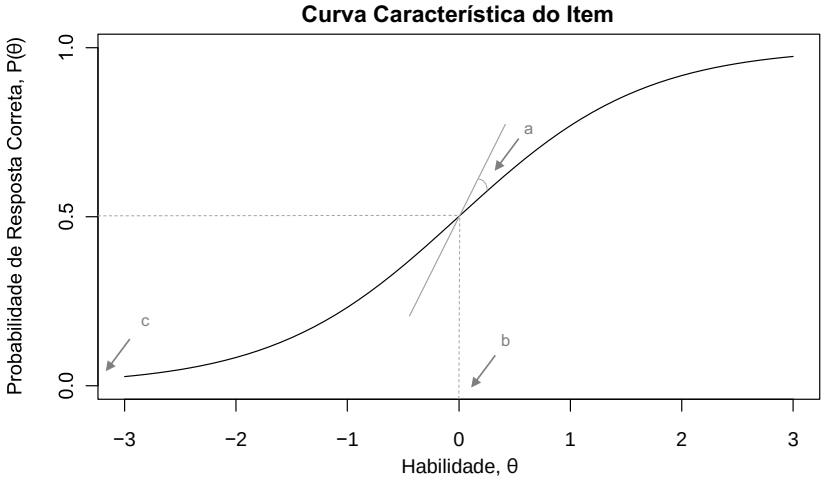

Figura 1: A Curva Característica do Item construída a partir dos parâmetros *a, b,* e *c*.

necessária para uma probabilidade de acerto igual a $\frac{(1+c)}{2}$, onde *c* independe da escala de habilidade, pois assume valores entre 0 e 1. Portanto, quanto maior o valor de *b*, mais difícil é o item, e vice-versa. Quando o modelo não leva em consideração o acerto casual, isto é, *c* assinalado com 0, a dificuldade *b* representa um ponto na escala de habilidade cuja probabilidade de resposta correta para um item *i* é de 50%. Quando esse fato ocorre, a CCI construída corresponde à probabilidade obtida do modelo Rasch.

O parâmetro *a* corresponde proporcionalmente ao ponto onde a reta tangente toca a curva. Valores baixos de *a* influenciam no poder de discriminação do item de modo que alunos com habilidades heterogêneas podem ter a mesma probabilidade de acertar um item. No entanto, Andrade et al. (2000) afirmam que o contrário, onde *a* assume valores muito elevados, indica uma curva bastante íngreme, logo o item tratado é considerado difícil, e portanto, é esperado que apenas indivíduos com habilidades elevadas consigam responder corretamente.

2.3.3 Estimação da Máxima Verossimilhança

Na Teoria da Resposta ao Item, a máxima verossimilhança é um procedimento aplicado na etapa de estimação das habilidades de um examinando. Uma vez que um modelo de TRI é selecionado com seus parâmetros e dados coletados, é possível avaliar, segundo Myung (2003), sua qualidade de ajuste, isto é, quão bem ele se ajusta aos dados observados. A qualidade do ajuste é avaliada encontrando os parâmetros do modelo de TRI que melhor se ajustam aos dados. O processo de

estimação inicia com o assinalamento de valores *a priori* para os parâmetros de habilidade θ e dificuldade do item e ocorre de maneira iterativa. A partir dessas estimativas, a probabilidade P(θ) é calculada para todos os itens assinalados por um indivíduo. O processo é repetido enquanto for possível obter ajustes do parâmetro de habilidade que melhorem a relação das probabilidades com os itens de resposta do examinando. A estimação é encerrada quando a diferença do ajuste se torna tão pequena que não demanda novas correções nas habilidades estimadas.

Conforme descreve Paek e Cole (2019), a estimação da máxima verossimilhança não fornece estimativas finitas podendo falhar em duas situações: o aluno acerta todos os itens do questionário e a habilidade estimada corresponde ao infinito positivo; ou o aluno não assinala alternativas corretas e a estimativa tende ao menos infinito. Nas situações apresentadas é impossível obter estimativas para o examinando. Em situações como essas, o recomendado é eliminar o aluno da análise.

A Subseção 2.4 descreve a etapa de calibração que visa encontrar as estimativas dos parâmetros a partir do conceito de máxima verossimilhança.

2.4 Teste de calibração

A técnica de calibração atua em conjunto com os modelos de TRI com a finalidade de estimar os parâmetros b e θ de um modelo que não são conhecidos *a priori* durante a avaliação de estudantes. O processo de calibração consiste em administrar um teste para N indivíduos que respondem a J itens de maneira dicotômica em um questionário. Após a aplicação do exame, alguns procedimentos são aplicados nas respostas, com a finalidade de expressar as habilidades identificadas em uma escala de habilidade comum entre os participantes do exame.

O Teste de Calibração definido para a execução dos experimentos deste artigo é baseado no paradigma de calibração introduzido por Allan Birnbaum em 1968 e implementado em (Baker e Kim, 2017). Birnbaum divide o paradigma em dois estágios de máxima verossimilhança. No primeiro estágio, a técnica estima os parâmetros de dificuldade dos J itens de um questionário. No segundo estágio, a partir dos parâmetros de dificuldade estimados, as habilidades dos N participantes são computadas. As estimações do segundo estágio são baseadas no conceito de ancoragem, o qual adota valores iniciais e realiza ajustes até encontrar estimativas que não demandem novos ajustes. O processo de estimação ocorre de modo iterativo buscando estimar os parâmetros que forneçam o melhor ajuste conforme descrito na Subsubseção 2.3.3. Detalhes específicos de implementação e tomada de decisões são explicados na etapa de experimentação da Subseção 4.3.

A Seção 3 descreve os trabalhos relacionados a esta pesquisa que utilizam a Teoria da Resposta ao Item em conjunto com as técnicas já apresentadas e que visam facilitar o processo de ensino-aprendizagem.

3 Trabalhos Correlatos

A descrição dos trabalhos relacionados foi obtida durante uma revisão *ad hoc* da literatura, com a finalidade de identificar pesquisas que lidassem com a Teoria da Reposta ao Item, aplicada a questionários de múltipla escolha, com o objetivo de fornecer *feedback* com base nas habilida-

des identificadas. Contudo, os trabalhos encontrados abordam os tópicos de maneira isolada ou a TRI com um dos tópicos mencionados. Entre os trabalhos encontrados há os de Chen e Duh (2008); Yarandi, Jahankhani, e Tawil (2013); Rajamani e Kathiravan (2013); El Falaki, El Faddouli, Idrissi, e Bennani (2013).

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. Para promover a eficácia da aprendizagem, um estudo anterior de Chen, Lee, e Chen (2005) propôs um sistema de *e-learning* personalizado baseado na teoria da resposta ao item (PEL-IRT), que pode considerar tanto a dificuldade de tópicos do curso quanto a capacidade do aluno avaliada pelas respostas de *feedback* (ou seja, entendimento completo ou não entendimento da resposta) para fornecer caminhos de aprendizagem personalizados para alunos de maneira individual.

Todavia, o PEL-IRT não pode estimar a capacidade do aluno para serviços de aprendizagem personalizados de acordo com as respostas não nítidas do aluno (ou seja, respostas incertas/imprecisas). O principal problema é que a resposta do aluno geralmente não pertence à compreensão completa ou à falta de entendimento do conteúdo dos cursos aprendidos. Portanto, o estudo apresenta um sistema de tutoria inteligente personalizado com base na Teoria da Resposta a Itens *Fuzzy* (FIRT), que pode ser capaz de recomendar cursos com níveis de dificuldade adequados para os alunos, de acordo com as respostas de *feedback* incerto ou *fuzzy* do aluno.

Em experimentos realizados para avaliar o desempenho da ferramenta, foram ministradas aulas de programação em Linguagem C com o tema "laço de repetição". Oitenta e oito alunos usaram a ferramenta e geraram um banco de dados de perfis de usuários com 1571 registros de aprendizado. Cada item respondido continha um parâmetro de dificuldade preestabelecido pelo professor. A escala de habilidade estava no intervalo $[-3, 3]$, ou seja, parâmetros de dificuldade com valores próximos a -3 indicavam uma questão fácil. Sempre que o aluno entra no sistema e não há registros próprios, a sua habilidade é inicializada em 0 o que indica uma habilidade moderada dada a escala de habilidade.

Os resultados obtidos indicam que a maioria dos alunos tem habilidades super-moderadas na unidade "Loop". O sistema proposto coleta as respostas dos alunos às perguntas: *"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. De acordo com os alunos conectados ao sistema na unidade "Loop", o grau médio de compreensão dos cursos recomendados é de 0,636, superior a 0,5, resultado que mostra que a compreensão dos alunos dos cursos recomendados é alta. Da perspectiva dos cursos, a proporção média dos cursos recomendados que podem ser compreendidos pelos alunos é de 0,627, acima de 0,5, resultado que também mostra que os alunos podem compreender os cursos mais recomendados na unidade "Loop".

Por sua vez, o trabalho de Yarandi et al. (2013) lida com sistemas de aprendizagem eletrônica (*e-learning*) apresentando uma abordagem diferente do que se tem pesquisado na literatura, pois os sistemas atuais adotam o conceito de aprendizagem estática, ou seja, os estudantes recebem o mesmo conteúdo independentemente do perfil que possuam. Assim, Yarandi et al. (2013) propõem um sistema de tomada de decisão que use 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, pois são a maneira mais comum de representar o conhecimento devido a sua flexibilidade e extensibilidade na modelagem de conceitos e seus relacionamentos, conforme explicam Esichaikul, Lamnoi, e Bechter (2011). Para cada aluno são geradas ontologias que descrevam o seu conhecimento com finalidade de representar um perfil de usuário ao qual são enviadas informações e conteúdos de aprendizagem compatíveis com as habilidades dos indivíduos. A identificação de habilidade adota a Teoria da Resposta ao Item como mecanismo de mensurar a habilidade de maneira acurada e, desse modo, os dados são coletados durante a interação do estudante com a ferramenta para continuamente serem fornecidos ao modelo, a fim de determinar as habilidades dos alunos de acordo com testes realizados no ambiente.

O sistema permite a aprendizagem adaptativa em dois aspectos:

1. Permite a apresentação de conteúdos para diferentes níveis de estudantes com características distintas.

2. 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.

Resultados da análise de questionário indicam que o sistema proposto melhorou a satisfação dos usuários, particularmente pelas capacidades adaptativas.

Permanecendo no campo de aprendizagem adaptativa, a pesquisa de Rajamani e Kathiravan (2013) gerou um sistema adaptativo de avaliação para compor testes digitais com o objetivo de melhorar a eficiência de elaboração de itens que se adéquem em múltiplos critérios de avaliação. O sistema proposto envolve duas fases:

1. Fase Preparatória: o professor seleciona parâmetros relacionados ao teste como variantes de conceitos a serem cobertos, relevância dos conceitos e critério de finalização. No começo do teste, geralmente são colocados itens previamente moderados com parâmetros predefinidos e a habilidade inicial do aluno é determinada antes do início do teste em si. Nesta proposta, um teste moderado é apresentado e, baseado nas respostas assinaladas, um teste sucessor é proposto. A abordagem dessa proposta adota um caminho diferente, ou seja, são aplicados testes de multiestágios, onde um conjunto de itens são apresentados e a habilidade é obtida com o início do teste propriamente dito.

2. Fase de Administração e Desenvolvimento do item: o teste pode ser conduzido de maneira adaptativa considerando todos os critérios de avaliação e todos os parâmetros são atualizados automaticamente. Dentre os critérios, cada item tem um valor ponderado para avaliar se o aluno compreendeu os conceitos, e cada pergunta está relacionada a um ou mais conceitos. O esquema de representação varia no intervalo [0, 4] conforme listado abaixo:

    - 0 – o item não tem relação com o conceito;

- 1 – o item tem fraca relação;
- 2 – o item é relacionado;
- 3 – o item tem alta relação;
- 4 – o item tem total relação.

A abordagem do uso de testes assistidos por computador é significativa e promissora na educação moderna, principalmente para liberar os professores do ônus de compor exames e melhorar a qualidade da avaliação dos testes. Os atributos da pergunta em um banco de perguntas são ajustados de forma adaptativa e dinâmica, sempre refletindo o status de aprendizado dos alunos. Várias outras tecnologias baseadas em inteligência artificial, ou otimização, podem ser exercitadas para desenvolver um teste mais eficiente, gerando abordagens para bancos de itens muito grandes. De acordo com Rajamani e Kathiravan (2013), a combinação de inteligência e personalização é a direção futura.

O último trabalho, de El Falaki et al. (2013), adota elementos da interação humano/computador, com a combinação cognitiva, comportamental e computacional, em um ambiente de aprendizagem. No escopo cognitivo e comportamental, a técnica de avaliação formativa adaptativa é empregada a fim de identificar o nível de competência do aluno e, a partir disso, fornecer orientações para que o aluno alcance o perfil de saída elaborado educacionalmente. No âmbito da computação, esse processo é realizado por meio de um sistema de *e-learning*, no qual a avaliação proposta é implementada por meio da arquitetura orientada a serviços (AOS).

Porém, diferente dos trabalhos citados, a pesquisa de El Falaki et al. (2013) oferece um sistema que individualiza o processo de avaliação, oferecendo um diagnóstico personalizado para decidir sobre a atividade de remediação. O foco está centrado na individualização do caminho de aprendizagem adotando a avaliação formativa. Assim, um teste adaptativo oferece uma seleção de itens ótimos em uma sequência, levando em consideração o perfil e o progresso do aluno.

Em um primeiro momento, aluno e itens são modelados de acordo com a abordagem baseada em competências. Em seguida, a avaliação formativa é modelada em uma abordagem adaptativa usando a teoria da resposta ao item, que visa fornecer uma série de itens selecionados consecutivamente. A resposta para um item determina a seleção do próximo, levando em consideração as respostas e os desempenhos anteriores registrados no modelo do aluno.

Por fim, a principal contribuição deste artigo é fornecer uma avaliação automática que construa *feedback* direcionado aos alunos e professores, com a finalidade de permitir aos alunos compreender de forma clara o porquê de suas dificuldades, expor a relação de suas habilidades com os tópicos compatíveis com o seu grau de habilidade e quais habilidades podem influenciar em uma melhora seu desempenho. Quanto aos professores, é possível ter uma visão geral do desempenho de uma turma destacando as questões mais fáceis, as mais difíceis, o acesso ao desempenho individual de cada estudante e a qualidade dos itens elaborados, tornando possível uma intervenção mais clara e específica do professor a fim de melhorar o processo de ensino-aprendizagem.

4 Metodologia

Esta seção detalha a metodologia empregada para a experimentação do modelo *Rasch* da Teoria da Resposta ao Item aplicado à base de dados coletada em uma escola amazonense de ensino médio. O objetivo do experimento é identificar as habilidades dos estudantes e dificuldades dos itens para gerar *feedback* formativo direcionado aos alunos e professores.

A Subseção 4.1 apresenta a estrutura dos dados coletados e a Subseção 4.2 detalha a etapa de tratamento destes dados de *log* para transformar o arquivo bruto em padrões que facilitem a extração de informação. Por fim, a Subseção 4.3 descreve a aplicação das técnicas e dos algoritmos utilizados na etapa de experimento.

4.1 Captura de dados

Para viabilizar a experimentação foi utilizada uma base de dados contendo o registro de vinte e um alunos de uma turma do segundo ano do ensino médio de uma escola pública do Amazonas. Estes alunos foram submetidos à avaliação de 5 disciplinas, resultando em 30 questões e produzindo um *log* bruto total de 983 registros.

Para complementar a base, outros dados como nome de disciplina, tópico em estudo e gabarito foram previamente gerados para auxiliar a correlação dos dados. A Tabela 1 exemplifica alguns dos dados extraídos do arquivo de *log* referente ao aluno com id_estudante "02".

Além dos atributos apresentados na Tabela 1, outros atributos necessários para a coleta e identificação de diferentes turmas estão inclusos na descrição abaixo:

- **id_estudante** – Identificação do aluno durante a realização de um simulado;
- **semestre** – Corresponde ao semestre letivo da coleta dos dados;
- **turma** – Código da turma participante do simulado;
- **tempo** – O valor corresponde ao *timestamp* em segundos, gerado no horário local com fuso GMT -4, referente ao momento de assinalamento de uma alternativa;
- **id_questao** – Identificador da questão;
- **id_questionario** – Identificador do questionário;
- **materia** – Nome da disciplina;
- **topico** – Tópico abordado dentro de uma disciplina;
- **alternativa** – Resposta do aluno para uma determinada questão.

A Subseção 4.2 detalha o processo de tratamento dos dados para a obtenção dos parâmetros de entrada do modelo *Rasch* a fim de viabilizar o início da etapa de experimento.

Tabela 1: Alguns dados contidos no *log* e correlacionados com a demais informações do primeiro semestre de 2019.

id_estudante	semestre	tempo (s)	id_questao	materia	topico	alternativa
02	201901	1573306013249	74	Física	Ondulatória	C
02	201901	1573306324030	75	Física	Ondulatória	B
02	201901	1573306434332	76	Física	Ondulatória	E
02	201901	1573306656113	77	Física	Ondulatória	A
02	201901	1573306859232	78	Física	Ondulatória	B
02	201901	1573309918932	82	Português	Leitura	B
02	201901	1573310107284	83	Português	Leitura de Textos Literários	A
02	201901	1573310109021	83	Português	Leitura de Textos Literários	B
02	201901	1573310109767	83	Português	Leitura de Textos Literários	A

4.2 Pré-processamento

Esta seção de tratamento dos dados tem a finalidade de remover ruídos da base de dados, produzir a correção dos questionários de cada disciplina, construir uma matriz binária de correção (baseado em certo ou errado) e gerar os parâmetros de entrada para o teste de calibração da TRI.

Os dados são copiados do *log*, que está no formato textual, para um banco de dados, com o intuito de facilitar a realização de consultas e manipulações dos dados. Os dados passam por um processo de limpeza e organização, principalmente para a retirada de dados duplicados e questões respondidas fora de ordem. Um exemplo de dados duplicados pode ser visto nas três últimas linhas da Tabela 1. Verificamos que estes dados correspondem à mesma questão, contudo as respostas e o tempo em que foram assinaladas divergem. Este caso em particular significa que o aluno respondeu a mesma questão três vezes. Portanto, a resposta para uma questão é válida como resposta final se o tempo de assinalamento for o maior entre todos os tempos para uma mesma questão. Para tratar o caso das questões fora de ordem é simplesmente feito uma ordenação.

Em seguida, a etapa de correção do questionário é iniciada. O objetivo é construir uma matriz binária indicando o erro ou acerto dos estudantes para cada questão de cada uma das disciplinas. As linhas da matriz correspondem às respostas de cada aluno e cada coluna contém 0 ou 1, conforme o aluno errou ou acertou a questão. A correção é baseada em uma consulta ao banco de dados *gabarito*.

A Tabela 2 apresenta a correção binária (erros ou acertos) para os 21 alunos da disciplina de Biologia. A tabela ainda contém a coluna de Pontuação Bruta que corresponde à soma total dos acertos de cada aluno.

Os parâmetros necessários para o teste de calibração derivam da manipulação dos dados contidos na Tabela 2. Todavia, devido a problemas com os critérios de convergência detalhados nos fundamentos da máxima verossimilhança (veja Subsubseção 2.3.3), tanto alunos que alcançaram a pontuação máxima, quanto alunos que não pontuaram são descartados da análise. De acordo com a Tabela 2 nenhum aluno acertou todos os itens, porém os alunos com identificação 13 e 16 obtiveram pontuação igual 0 e, portanto, seus registros são removidos.

Tabela 2: Respostas dos alunos para a disciplina de Biologia.

Aluno	**Item**					**Pontuação Bruta**
	170	171	172	173	174	
01	1	1	0	1	1	4
02	0	0	0	1	0	1
03	0	0	1	0	1	2
04	0	1	0	1	1	3
05	1	1	1	1	0	4
06	0	0	0	1	0	1
07	0	0	1	1	1	3
08	0	0	1	1	0	2
09	1	1	0	1	1	4
10	0	0	1	0	1	2
11	1	1	0	1	0	3
12	0	0	0	1	0	1
13	0	0	0	0	0	0
14	1	0	1	1	1	4
15	0	0	1	1	1	3
16	0	0	0	0	0	0
17	0	0	0	1	0	1
18	0	1	1	1	0	3
19	0	0	0	1	1	2
20	0	0	1	1	0	2
21	0	1	0	1	1	3

Para a geração dos dois vetores de parâmetros necessários para a etapa de experimentação, a metodologia impõe que, a partir da Tabela 2, seja gerado uma nova tabela (Tabela 3), denominada tabela de frequência, e que conterá os dois vetores de parâmetros. É possível verificar que a Tabela 3 contém a coluna *Pontuação* com *N* linhas, sendo que, os valores nesta coluna dependem da pontuação mínima e máxima, obviamente observando que *tanto alunos que alcançaram a pontuação máxima quanto alunos que não pontuaram são descartados da análise*. No caso da Tabela 2, o valor da *Pontuação* varia de 1 até 4, indicando que os alunos acertaram 1, 2, 3 ou 4 questões, formando, nesse caso específico, quatro conjuntos de alunos. O grupo 1 (que acertaram somente uma questão) é composto pelos alunos “02”, “06”, “12” e “17”. O grupo 2 (que acertaram duas questões) pelos alunos “03”, “08”, “10”, “19” e “20”. O grupo 3 (que acertaram três questões) pelos alunos “04”, “07”, “11”, “15”, “18” e “21”. E o grupo 4 (que acertaram quatro questões) pelos alunos “01”, “05”, “09” e “14”. A partir de cada grupo soma-se as quantidades de acertos em cada questão. No caso do grupo 1, por exemplo, todos os quatro alunos só acertaram a questão 173 e erraram as outras. No caso do grupo 2, os cinco alunos tiveram 4 acertos na questão 172, 3 acertos na questão 173 e 3 acertos na questão 174. Os outros dois grupos são feitos dessa mesma forma. O total de acertos em cada grupo pode ser visto na coluna *Soma das Linhas* da Tabela 3. A coluna *Frequência da Pontuação* é obtida pela divisão da coluna *Soma das Linhas* pela coluna *Pontuação*; e a linha *Soma Itens* corresponde à soma total de cada uma das colunas da Tabela 3.

Por fim, os parâmetros do teste de calibração são determinados pelos valores da linha *Soma Itens*, ou seja, $(5,7,9,17,10)$; e os valores da coluna *Frequência de Pontuação*, ou seja, $(4,5,6,4)$. Vale a pena mencionar que a linha *Soma Itens* está relacionada com as colunas *Item* e a coluna *Frequência da Pontuação* está relacionada com as linhas *Pontuação*. A etapa de pré-processamento é encerrada com formação da tabela de frequências para todas as disciplinas. A Subseção 4.3 detalha a etapa de experimentação dos dados submetidos aos algoritmos da TRI durante o teste de calibração para a identificação das habilidades dos estudantes.

Tabela 3: Contagem das frequências para as respostas dos alunos de Biologia.

Pontuação	**Item**					**Soma das Linhas**	**Frequência da Pontuação**
	170	**171**	**172**	**173**	**174**		
1	0	0	0	4	0	4	4
2	0	0	4	3	3	10	5
3	1	4	3	6	4	18	6
4	4	3	2	4	3	16	4
Soma Itens	5	7	9	17	10	48	19

4.3 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 técnica de Teste de Calibração, definida na Subseção 2.4, é utilizada para auxiliar nesta finalidade. O experimento segue o paradigma de Birnbaum, pois utiliza o modelo Rasch, que trabalha bem com uma quantidade pequena de itens e estudantes. 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 e a soma dos itens, vetores que foram obtidos durante a etapa de pré-processamento (veja a Subseção 4.2).

O experimento foi executado para as disciplinas Biologia, Física, Língua Inglesa, Língua Portuguesa e Química. Destas, quatro disciplinas tinham 5 questões com exceção de Língua Portuguesa que continha 10 questões. Contudo, os dados das disciplinas Física, Língua Inglesa e Química apresentaram resultados muito semelhantes em termos de informação. Para evitar repetição de informação, os resultados analisados são baseados nas disciplinas de Biologia e Língua Portuguesa que apresentaram resultados significativos para serem apresentados. Para exemplificar o processo de calibração, a análise dos dados continua a partir da Tabela 3 correspondente à disciplina de Biologia. O atributo *Soma Itens* passa a ser definido pelo vetor $s = (5,7,9,17,10)$, e a *Frequência da Pontuação* passa a ser definida pelo vetor $f = (4,5,6,4)$.

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 Equação 3.

$$b = \left(\log \left(\frac{\sum_{n=1}^{|f|}(f_n) - s_1}{s_1} \right), ..., \log \left(\frac{\sum_{n=1}^{|f|}(f_n) - s_j}{s_j} \right) \right), \tag{3}$$

onde $|f|$ corresponde ao total de elementos em f e j coincide com o valor de J.

O parâmetro de habilidade θ é definido, em um primeiro momento, 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 Equação 4.

$$\theta = \left(\log \left(\frac{1}{J-1} \right), ..., \log \left(\frac{N}{J-N} \right) \right), \text{onde N é o total de elementos em } \theta. \tag{4}$$

Por fim, resta definir um ponto médio para a escala de habilidade, onde o parâmetro de dificuldade do item atua como valor de localização, isto é, indica a habilidade esperada para o aluno assinalar o item corretamente. Levando em consideração a implementação do programa BICAL de Wright e Mead (1980) onde o ponto médio corresponde à média dos elementos no vetor b, a implementação deste experimento, conforme descrito no Algoritmo 1 baseado em Baker e Kim (2017), utiliza o mesmo critério como referência, porém os parâmetros em b são ajustados a partir da subtração de b pela sua média $\overline{b}$ de modo que o ponto médio dos itens ajustados seja 0. Esse ajuste facilita a localização e posterior interpretação de uma habilidade na escala obtida. Por exemplo, em uma escala que varia de [-3,3], um aluno com habilidade 0 tem 50% de chances de acertar um item i com dificuldade moderada (b = 0). Um aluno com habilidade 2 tem maior probabilidade de acertar uma questão do que um aluno teria com habilidade -3. O Algoritmo 1 apresenta a função principal do teste calibração. O teste é dividido em dois estágios seguindo o paradigma de Birnbaum. Estes estágios ocorrem em um laço de repetição com k ciclos de iteração que podem ser interrompidos se a checagem de convergência atender ao critério de parada e, neste caso, as habilidades estimadas satisfazem a calibração e a métrica da escala de habilidade é encontrada.

O primeiro estágio recebe como entrada os parâmetros b, θ, s e f. O objetivo deste estágio é determinar o parâmetro de dificuldade do item. Ao final do estágio um, o parâmetro b é atualizado com os valores computados. O Algoritmo 2 descreve as etapas do primeiro estágio para a estimação dos parâmetros de dificuldade do vetor b correspondentes aos itens avaliados. O laço mais interno, linhas 7 - 11, computa a probabilidade de um aluno com habilidade θ_g acertar uma questão com dificuldade b_j e assinala o valor à variável p. O valor da probabilidade é multiplicado pelo valor da frequência f_g e salvo na variável acumuladora *somafp*. Ainda nesse bloco de repetição, a variável acumuladora *somafpq* recebe a mesma multiplicação do passo anterior com a inclusão de um termo na conta: $1 - p$ que representa a probabilidade de um aluno não acertar a questão. Os dois acumuladores são utilizados para obter a variação *deltab* e atualizar parâmetros de b a partir de sua variação em uma operação de subtração. Ao final do primeiro estágio é verificado se o valor absoluto *deltab* é menor do que a constante *convb* referente ao critério de convergência. Se o critério não for satisfeito os cálculos são refeitos a partir dos valores atualizados para b; caso contrário o laço é interrompido. Isso significa que a soma da diferença entre b e a variação *deltab* se torna tão pequena que uma próxima iteração não altera os valores da variação, logo o valor estimado de b é usado como parâmetro de dificuldade.

Algoritmo 1 Função de calibração para o modelo Rasch

```
1: calibracaoRasch(s, f) {
2: J ← tamanho(s); G ← tamanho(f); K ← 25
3: T ← 10; b ← log (soma(f) − s) / s
4: b ← b − media(b)
5: bAntigo ← b; θ ← sequencia(1, G, 1)
6: para g = 1 até G faça
7:     θ_g ← log g / (J − g)
8: fim para
9: para k = 1 até K faça
10:    convabd ← 0,01
11:    b ← estagioUm(b, θ, s, f)
12:    b ← b − media(b)
13:    θ ← estagioDois(θ, b)
14:    abd ← abs(b − bAntigo)
15:    se soma(abd) < convabd então
16:        interromper
17:    senão
18:        bAntigo ← b
19:    fim se
20: fim para
21: b ← b × (J − 1) / J
22: θ ← estagioDois(θ, b)
23: θ ← θ × (J − 2) / (J − 1)
24: retorna (θ, b)}
```

Algoritmo 3 Estimação do parâmetro θ a partir dos valores de b.

```
1: estagioDois (θ, b) {
2: G ← tamanho(θ); J ← tamanho(b); T ← 10
3: para g = 1 até G faça
4:     convt ← 0,01
5:     para t = 1 até T faça
6:         somap ← 0; somapq ← 0
7:         para j = 1 até J faça
8:             p ← 1 / (1 + e^−(θ_g − b_j))
9:             somap ← somap + p; somapq ← somapq − p × (1 − p)
10:        fim para
11:        deltat ← (g − somap) / somapq
12:        θ_g ← θ_g − deltat
13:        se abs(deltat) < convt então
14:            interromper
15:        fim se
16:    fim para
17: fim para
18: retorna θ
19: }
```

Algoritmo 2 Função do primeiro estágio da calibração para estimar o parâmetro b.

estagioUm (b, θ, s, f) {
$J \leftarrow$ **tamanho**(b); $G \leftarrow$ **tamanho**(θ); $T \leftarrow 10$
para $j = 1$ **até** J **faça**
  $convb \leftarrow 0,01$
  para $t = 1$ **até** T **faça**
    $somafp \leftarrow 0$; $somafpq \leftarrow 0$
    para $g = 1$ **até** G **faça**
      $p \leftarrow \frac{1}{1+e^{-(\theta_g - b_j)}}$
      $somafp \leftarrow somafp + f_g \times p$
      $somafpq \leftarrow somafpq + f_g \times p \times (1-p)$
    fim para
    $deltab \leftarrow \frac{s_j - somafp}{somafpq}$
    $b_j \leftarrow b_j - deltab$
    se abs$(deltab) < convb$ **então**
      interromper
    fim se
  fim para
fim para
retorna b
}

Os cálculos do segundo estágio são computados de maneira semelhante. O objetivo é obter o vetor de habilidades a partir dos parâmetros de dificuldade já estipulados. A variação de θ e *deltat* é obtida a partir dos acumuladores *somafp* e *somafpq*. Em seguida, o parâmetro θ é atualizado a partir de *deltat* conforme descrito no Algoritmo 3. Por fim, o critério de convergência verifica se o valor absoluto de *deltat* é menor do que a constante *convt*. Se o critério de parada for satisfeito, então θ é retornado, caso contrário, a estimação continua a partir do valor atual de θ.

5 Resultados e Discussões

5.1 Análise e Correlação das Informações Obtidas da TRI

Tabela 4: Parâmetros de dificuldade para a disciplina de Biologia.

Item	Dificuldade	Tópicos
173	-2,1234	Divisão Anatômica
174	0,0003	Anatomia do Sistema Nervoso
172	0,2304	Sistema Nervoso
171	0,6949	Ação Hormonal
170	1,1982	Sistema Endócrino

A Tabela 4 elenca os parâmetros de dificuldade estimados para os itens da disciplina Biologia. O item 170, referente ao tópico de *sistema endócrino*, corresponde à questão mais difícil da disciplina. Seu parâmetro de dificuldade está estimado em $1,1982$. Ao correlacionar esse valor

com a quantidade de acertos do item na Tabela 3, é possível compreender que a estimativa está de acordo com o total acertos, pois dos 19 alunos apenas 5 responderam corretamente. Associando essa linha de raciocínio com a curva característica correspondente a este item na Figura 2, e observando as habilidades estimadas da Tabela 5, é possível compreender que, para esta população, apenas alunos com habilidade igual ou superior a $1,28$ têm maior probabilidade de assinalar o item corretamente. Este fato é justificado, pois o parâmetro de dificuldade é um valor que reside na escala de habilidade. Assim, se $b_i = \theta_j$ e b_i = 1,1982, então esta habilidade seria a ideal para o aluno ter probabilidade de 50% para assinalar o item 170 corretamente. Todavia, entre as habilidades estimadas na Tabela 5, apenas a habilidade de 1,28 está próxima da dificuldade 1,1982 e, portanto, apenas os alunos 1, 5, 9 e 13 do grupo 4 têm mais chances de assinalar o item 170. Como 1,28 corresponde a maior habilidade identificada, o mesmo vale para os demais itens de Biologia. Em termos práticos, a Equação 5 do modelo Rasch ajuda a entender as chances de acerto.

$$\begin{aligned} P(U_{ij} = 1|\theta_j) &= \frac{1}{1+e^{-1(1,28-1,1982)}} \\ &= \frac{1}{1+e^{-0,081}} \\ &= 0,5204 \end{aligned} \tag{5}$$

Ao atribuir os valores de 1,28 e 1,1982 para θ_j e b_i, respectivamente, a probabilidade para um aluno do grupo 4 assinalar o item corretamente é de 52,04%. De outra forma, se a habilidade θ for substituída pela estimativa imediatamente inferior, 0,45, então a probabilidade é de apenas 32,12%. Logo, é possível afirmar que 1,28 contribui para uma maior probabilidade de acerto.

Tabela 5: Habilidades obtidas para o teste de Biologia.

Pontuação Bruta	Frequência da Pontuação	Habilidade	Identificador do Aluno
1	4	-1,30	02, 06, 12, 17
2	5	-0,31	03, 08, 10, 19, 20
3	6	0,45	04, 07, 11, 15, 18, 21
4	4	1,28	01, 05, 09, 14

Entretanto, o aluno de identificador "11" é uma exceção, pois assinalou o item corretamente estando fora desse grupo, sua habilidade é de $0,45$, e a probabilidade de acerto deste aluno é de $32,12\%$, cerca de $19,92\%$ abaixo da probabilidade mínima de acerto, o que sugere uma investigação do fato.

Por outro lado, analisando o item 173, que apresenta a maior discrepância em relação aos demais itens em termos de CCI, é possível notar pelo gráfico que muitos alunos obtiveram êxito no tópico de *divisão anatômica*, o que permite concluir que este seja o item mais fácil do questionário e que este fato impactou diretamente em uma menor estimativa de dificuldade, $-2,1234$.

A Tabela 3 reforça que 17 alunos acertaram o item e apenas os alunos "03" e "10", ambos com habilidade $-0,31$, falharam na questão mesmo com uma probabilidade de acerto estimada em 86%. Aqui cabe um ponto de investigação. Por ser uma questão fácil todos os alunos com a menor habilidade identificada, $-1,30$, responderam corretamente o item, o que era totalmente esperado.

A Tabela 6 apresenta as habilidades estimadas para os alunos da disciplina de Língua Portuguesa e Literatura. Para este questionário, o total assinalamentos corretos está no intervalo $[3,7]$

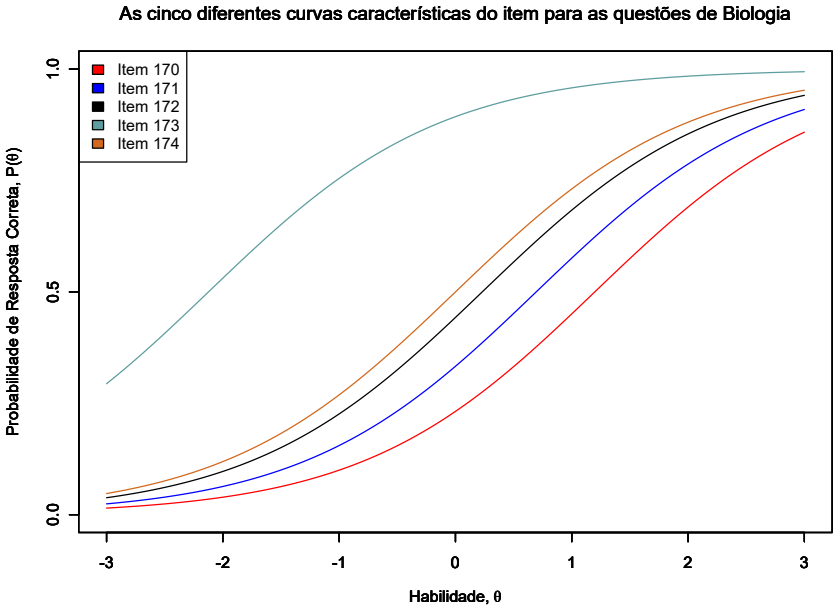

Figura 2: Curvas Características para as cinco questões de Biologia.

de modo que nenhum aluno foi agrupado em um conjunto com as demais habilidades. Este fato justifica a frequência assinalada com 0 em algumas linhas da tabela.

Tabela 6: Habilidades obtidas para o teste de Língua Portuguesa.

Pontuação Bruta	**Frequência da Pontuação**	**Habilidade**	**Identificador do Aluno**
1	0	-2,7109	-
2	0	-1,7186	-
3	3	-1,0157	02, 04, 08
4	7	-0,4398	03, 05, 12, 13, 14, 17, 19
5	7	0,07853	01, 06, 09, 16, 18, 20, 21
6	3	0,5824	07, 10, 15
7	1	1,1093	11
8	0	1,7149	-
9	0	2,5542	-

Como é possível notar na Tabela 7, a quantidade de itens influencia proporcionalmente em uma maior quantidade de dificuldades observadas para uma mesma disciplina. Os dados apresentados estão ordenados de forma crescente a partir do atributo dificuldade.

O item 93 é o mais fácil deste questionário, pois alunos com a menor habilidade estimada (-1,0157) têm probabilidade de $87,75\%$, conforme cálculos da Equação 6, para assinalar correta-

Tabela 7: Parâmetros de dificuldade para a disciplina de Língua Portuguesa.

Item	Dificuldade	Tópicos
93	-2,9845	Literatura
92	-1,8668	Leitura
87	-0,6068	Leitura do Implícitos
83	-0,4119	Leitura de Textos Literários
88	-0,0332	Leitura
85	0,1572	Leitura Brasileira
82	0,7764	Leitura
95	1,2917	Elementos de Coesão
97	1,6219	Literatura
94	2,0560	Leitura Brasileira

mente o tópico abordado em Literatura.

$$\begin{aligned} P(U_{ij}=1|\theta_j) &= \frac{1}{1+e^{-1(-1,0157-(-2,9845))}} \\ &= \frac{1}{1+e^{1.9688}} \\ &= 0.8775 \end{aligned} \tag{6}$$

Em contrapartida, o item 94 sobre leitura brasileira é o mais difícil, com estimativa de 2,0560. A Equação 7 demonstra que o aluno com a maior habilidade estimada, 1,1093, tem apenas 27,95% de probabilidade de acerto. Analisando a Figura 3 fica nítida a diferença de dificuldade entre as questões.

$$\begin{aligned} P(U_{ij}=1|\theta_j) &= \frac{1}{1+e^{-1(1,1093-2,0560)}} \\ &= \frac{1}{1+e^{0.9467}} \\ &= 0.2795 \end{aligned} \tag{7}$$

Para as três questões mais difíceis da disciplina, o cálculo das probabilidades segue os princípios da Equação 6 e Equação 7. A probabilidade que mais se aproxima de 50%, levando em consideração a habilidade mais elevada, é a da questão 95 com 45,45%. A partir da questão 82 a probabilidade cresce e ultrapassa os 50% tendo valor igual a 58,25% e continua a crescer, dado que a dificuldade tende a diminuir e as curvas se distanciarem da questão 94.

A partir das informações obtidas, artefatos de *feedback* automático foram construídos para fornecer aos alunos e professores uma melhor compreensão das informações.

5.2 Análise do *Feedback* Automático

Um protótipo do ambiente de *feedback*[4] foi implementado através de páginas web, separadas por disciplina, as quais apenas o aluno e professor têm acesso mediante autenticação.

[4]https://simulado-rbie-v1.herokuapp.com/

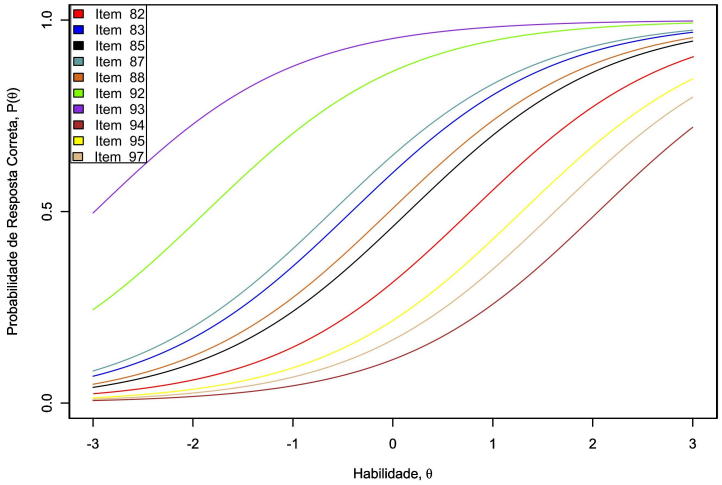

Figura 3: Curvas Características para os itens de Língua Portuguesa.

O *feedback* direcionado ao aluno contém as curvas características de cada item para uma determinada disciplina, além de suas habilidades, probabilidade de acerto, dificuldade da questão e o grupo ao qual o aluno pertence com base em seu padrão de respostas. Por exemplo, a Figura 4 apresenta *feedback* para o item 170 de Biologia do aluno de id 04. Este aluno foi classificado no grupo de pontuação 3, ou seja, do total de itens assinalados apenas 3 estavam corretos, conforme apresentado na Tabela 5. A habilidade estimada para o aluno foi de 0,45 o que produz uma probabilidade de acerto em torno de 32,02% conforme indicado na CCI. Contudo, a habilidade identificada é inferior a dificuldade do item, cerca de 0,75 de diferença, constatando que o aluno tinha poucas chances de acertar o tópico de Sistema Endócrino. Portanto, o *feedback* deve sugerir ao aluno uma maior dedicação para aumentar as suas habilidades especificamente no tópico de Sistema Endócrino.

Os artefatos construídos pela Teoria da Resposta ao Item e fornecidos como *feedback* elaborado ao professor contribuem não apenas com informações sobre o desempenho de estudantes, mas também com subsídios para que o próprio docente analise se os itens elaborados seguem o propósito da avaliação para o qual foram designados, isto é, se os itens de um exame cumprem o papel de distinguir o nível de habilidade dos discentes e também permitir a identificação de eventuais equívocos na metodologia. Por exemplo, ao receber as informações sobre o desempenho de uma turma, o professor pode verificar o nível de dificuldade estimado para um item ao analisar curva característica. Se probabilidade de acerto está muito elevada para habilidades medianas (abaixo de zero), então o item foi considerado fácil pelos examinandos. Nesta situação, portanto, o professor pode compreender que:

- Como o item é fácil, este foi elaborado corretamente e não deve ser aplicado quando o objetivo for fazer a distinção entre os alunos com habilidades elevadas e alunos com habilidades medianas; ou
- O professor pode constatar que cometeu um equívoco ao elaborar o item, pois a intenção era compor um item de dificuldade elevada. Se for esse o caso, a informação da dificuldade do item deve ser corrigida tão logo quanto possível, até mesmo para evitar retrabalhos do professor.

Analisando outro cenário, no qual muitos alunos erraram um item, a curva característica deve refletir com precisão que a questão avaliada é difícil. Portanto, o professor pode assimilar algumas possibilidades:

- O item é indicado para aplicar à uma classe quando o objetivo é identificar quais os alunos possuem habilidades elevadas; ou
- O professor pode ter cometido um equívoco na definição do grau de dificuldade, imaginado que a questão seria mais fácil e, portanto, a definição do item deve ser ajustada; ou
- Houve uma falha de comunicação no processo de ensino-aprendizagem e, assim, o professor pode rever se há a necessidade de reforço no tópico abordado.

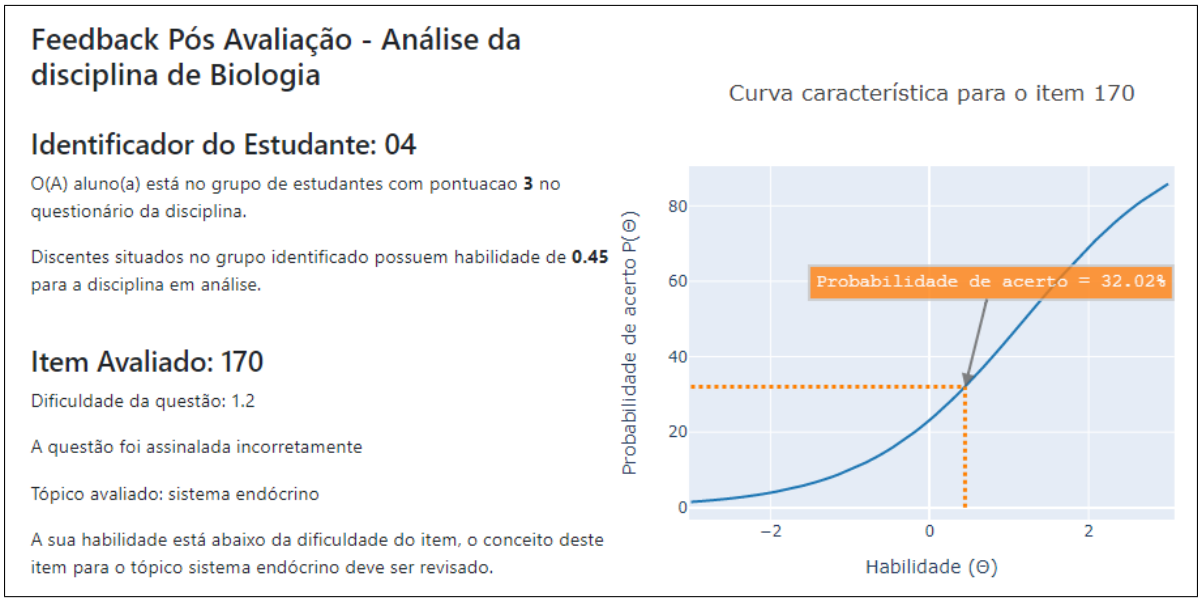

Figura 4: Trecho do relatório de *feedback* para o aluno de id 04 abordando a disciplina de Biologia.

O *feedback* fornecido ao professor é semelhante ao destinado para o aluno com o acréscimo das tabelas geradas durante o experimento e analisadas na Subseção 5.1, como a Tabela 2 de assinalamentos, a Tabela 4 de dificuldade dos itens e os gráficos, como o da Figura 3, que contém todas curvas características para uma determinada disciplina.

Os artefatos elencados nesta seção como tabelas e gráficos, dispensam a necessidade do professor aplicar diretamente a TRI em suas avaliações, pois o *feedback* automático sumariza os artefatos em páginas web por disciplina, o que favorece o uso do *feedback* 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 em caráter de múltipla escolha e opta por um método de correção automática dos itens que forneça *feedback* elaborado.

A Subsubseção 5.2.1 detalha a experimentação e as percepções de professores sobre o ambiente de *feedback* elaborado a partir da base de dados em estudo neste trabalho.

5.2.1 Validação do Protótipo pelo Professor

Para validar o método de *feedback* automático junto a usuários, professores que originalmente elaboraram as questões de disciplinas como Física, Língua Inglesa, Língua Portuguesa e Química avaliaram os artefatos de *feedback* disponibilizados em páginas web conforme exemplifica a Figura 4. Diante das restrições impostas pela pandemia de covid-19, como escolas sem aulas presenciais ou até ociosas, além de outros serviços interrompidos, o grupo foi composto apenas por professores (quatro no total), com conhecimentos moderados sobre o uso de computadores de maneira geral. Todos os professores participaram de um treinamento *on-line* assíncrono por meio de material elaborado para instruí-los sobre os conceitos básicos da Teoria da Resposta ao Item e utilização do ambiente, de modo que cada professor estivesse apto a analisar o protótipo de *feedback*.

Os professores responderam ao questionário pós-experimento que tinha a finalidade de identificar tanto as percepções quanto o uso do ambiente. Para este contexto, adaptou-se o modelo de avaliação focado em aceitação e uso de ambientes para aprendizagem eletrônica destinado aos professores e alunos proposto por Umrani-Khan e Iyer (2009). Os professores avaliaram o ambiente levando em consideração afirmativas associadas aos aspectos de: (a) utilidade percebida, (b) interatividade e (c) facilidade de uso. Para assinalar a concordância das afirmativas relacionadas a cada aspecto em avaliação foi utilizada a escala *likert* contendo 5 pontos: discordo totalmente, discordo parcialmente, neutro, concordo parcialmente, e concordo totalmente. A Figura 5 contém o assinalamento paras as afirmativas de utilidade e interatividade e a Figura 6 destaca o assinalamento para as afirmativas quanto a facilidade de uso.

Além da análise em escala likert foram aplicadas perguntas abertas: a) *"Qual é sua opinião quanto as informações de avaliação e feedback automáticos do ambiente para auxiliar na interpretação de possíveis dificuldades/dúvidas dos alunos de sua turma?"*; b) *"Você acha útil a interatividade que o gráfico proporciona para o usuário seja ele discente/docente?"*; c) *"Sobre o gráfico, pode-se simular qual a probabilidade de acerto alterando a habilidade do aluno. Você acha que tal opção pode motivar o aluno a querer aumentar suas habilidades?"*; d) *"Você achou útil a apresentação dos dados em tabelas?"*; e) *"O detalhamento das habilidades dos alunos te ajuda a fazer um acompanhamento individual ?"*; f) *"As dificuldades identificadas e organizadas em tabela te ajudam a saber quais medidas dever ser tomadas para melhorar o rendimento da turma?"*; g) *"Ao elaborar questões de um tópico você planeja o nível de dificuldade e facilidade. Contudo, ao receber o feedback do ambiente você descobre que alguma questão foi identificada como fácil ou difícil pela turma quando a sua intenção era o oposto. Esse feedback te ajuda a repensar se a questão foi elaborada corretamente? Se deve ser reescrita? Ou se algo na metodologia precisa ser corrigido?"*; h) *"O que dificultou o uso da ferramenta de apoio para analisar o desempenho dos alunos?"*; i) *"O que você mudaria no uso da ferramenta de apoio para melhorar a sua compreensão?"*; j) *"O que você mudaria no uso da ferramenta de apoio para melhorar a forma como as informações são apresentadas?"*.

O resultado obtido do experimento para a avaliação dos artefatos de *feedback* permite concluir pela Figura 5 que em termos de utilidade e interatividade a divisão entre escolhas foi satisfatória estando entre a concordância total ou concordância parcial. Os termos *"Eu acho o ambiente útil no meu ensino"* e *"Usando o ambiente, posso interagir com os alunos e esclarecer suas dúvidas em um tempo razoável"* tiveram a concordância total dos quatro professores. Para as demais

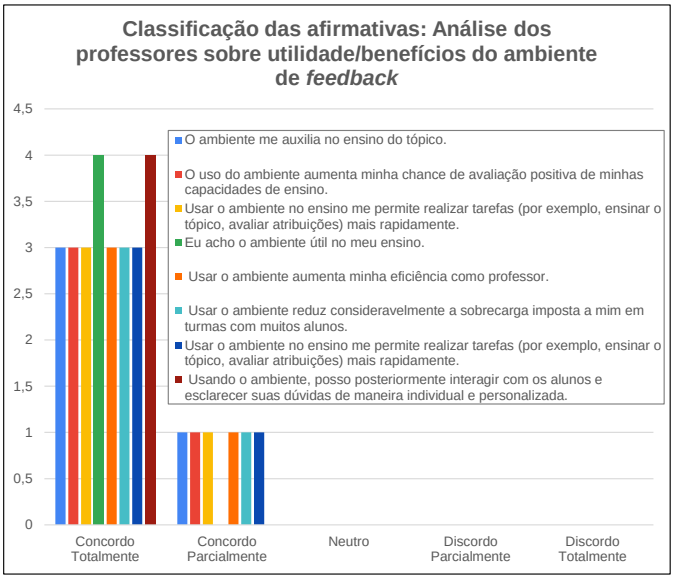

Figura 5: Resultados sobre utilidade/benefícios do *feedback*.

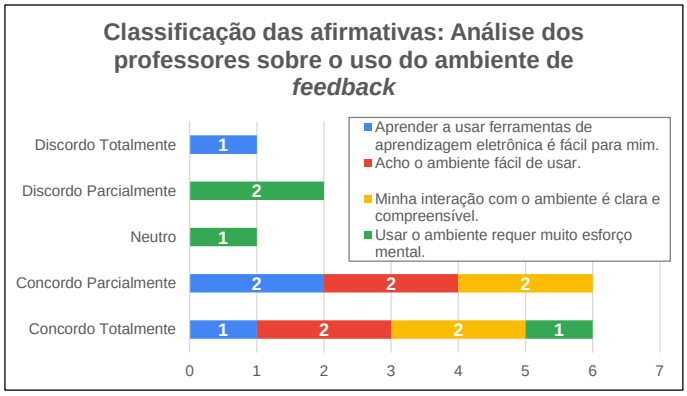

Figura 6: Resultados sobre o uso dos artefatos de *feedback*.

afirmativas três professores concordaram totalmente e um concordou parcialmente. De maneira geral, foram 26 concordâncias totais e 6 concordâncias parciais. A partir da análise é possível compreender que os professores julgam possível o aumento da eficiência ao lecionar utilizando o ambiente e que isso é proporcionado pela diminuição de sobrecarga ocasionada pelo *feedback*. Por fim, os docentes concordam que o *feedback* pode auxiliar na interação e acompanhamento individual e personalizado para cada aluno.

A análise levando em consideração a necessidade de um esforço mental para usar o ambiente, conforme ilustra a Figura 6, demonstra que dois professores discordam parcialmente quanto a afirmativa *"Usar o ambiente requer muito esforço mental"*. A dificuldade de alguns professores pode ser entendida devido ao contato em um intervalo curto de tempo com a TRI e o fato do treinamento ser assíncrono por motivos de distanciamento social e aulas presenciais suspensas dada a situação de saúde no Brasil. Pode-se notar que nem todos os professores têm domínio de ambientes virtuais para a aprendizagem, pois um discorda totalmente e outros dois concordam parcialmente, ou seja, é necessário um treinamento mais abrangente sobre o uso dos artefatos do

ambiente. Por fim, as afirmativas sobre facilidade de uso e interação clara e compreensiva obtiveram duas concordâncias totais e duas parciais. Apesar de utilizar uma teoria mais complexa do que a TCM, os professores conseguem entender o feedback elaborado e seus objetivos para auxiliar o processo de ensino-aprendizagem. Em linhas gerais, a avaliação apresentada na Figura 6 segue a tendência de maior concordância com as afirmativas apresentadas dado que houve apenas uma discordância total e duas parciais de modo que os artefatos de *feedback* elaborados em páginas web receberam boa aceitação dos professores em sua maioria.

Para entender os aspectos que impactaram de maneira positiva ou negativa a análise dos professores, foram analisadas as respostas das perguntas abertas do formulário para identificar a razão das concordâncias e discordâncias com os aspectos avaliados. A seguir são apresentados alguns comentários positivos e negativos considerando as respostas, onde P_i corresponde ao i-ésimo participante:
"Acredito que são muito boas mas sendo que será um necessário esclarecimento ou fica como dica a criação de um vídeo tutorial explicando alguns dados, vídeos pequenos, ajudariam a entender." − Negativo (P3) sobre a pergunta a).
"Otimizam o tempo do professor, além de dar uma visão detalhada das deficiências encontradas na assimilação do conteúdo." − Positivo (P4) sobre a pergunta a).
"Sim, pois o o tempo que eu gastaria nas correções eu posso utilizar na pesquisa ou criação de material." − Positivo (P4) sobre a pergunta e).
"Deixa nosso trabalho mais ágil, e mostra o grau de compreensão do tópico avaliado." − Positivo (P1) sobre a pergunta d).
"Acredito que este aspecto está bom e não precisa de mudanças muito extremas, só melhorar o layout para ficar mais friendly." − Negativo (P02) sobre a pergunta j).
"Poderíamos ter algumas videoaulas gravadas com explicação de como utilizar a ferramenta, isso possibilitaria seu uso em várias partes do país" − Negativo (P1) sobre a pergunta j).
"Sim, contribui em poder visualizar isso junto com os alunos ou direcionar o docente nas ações a seguir." − Positivo (P3) sobre a pergunta d).

6 Conclusão

O uso da Teoria da Resposta ao Item 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 turma de estudantes. O *feedback* elaborado proporciona uma comunicação mais direta e objetiva entre alunos e professores por meio de acesso simples, didático e interativo às informações obtidas a partir da análise de rendimento dos estudantes. Portanto, os artefatos de *feedback* construídos com base nas estatísticas geradas pela TRI, e sumarizados em páginas web, 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 o tópico avaliado também permite o estudo de conteúdo compatível com sua habilidade.

Do ponto de vista do professor, as respostas obtidas na experimentação permitem concluir que as informações apresentadas otimizam o 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, como questões mal formuladas, com o propósito de adequar seus instrumentos de avaliação a cada turma.

Para trabalhos futuros, há o interesse de acrescentar outros modelos logísticos da TRI, isto é, com dois e três parâmetros, os quais podem fornecer mais informações tais como a discriminação de itens e o acerto casual (chute), conforme descritos na Subsubseção 2.3.1. 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 e os que não têm a possibilidade, além de ilustrar e penalizar respostas por acerto casual. Além dos modelos, as repostas coletadas pós-avaliação dos professores indicam a necessidade de incluir mais detalhes sobre o uso do protótipo, podendo ser um vídeo ou um manual detalhado. Pretende-se evoluir o desenvolvimento do método para torná-lo público, assim professores podem acessar o ambiente e inserir suas próprias bases de dados para que o modelo construa os artefatos de *feedback* automático. Em termos visuais, diante da sugestão coletada na experimentação do protótipo, o ambiente de *feedback* pode ser melhorado com o uso de algum *framework* para o desenvolvimento da interface como o *bootstrap*.

Agradecimentos

Esta pesquisa, conforme previsto no Art. 48 do decreto nº 6.008/2006, foi parcialmente financiada pela Samsung Eletrônica da Amazônia Ltda, nos termos da Lei Federal nº 8.387/1991, através de convênio nº 003/2019, firmado com o ICOMP/UFAM. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001. Este trabalho também contou com o apoio financeiro parcial da Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), por meio dos projetos 122/2018 (UNIVERSAL) e 003/2019 (POSGRAD).

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*. [GS Search]

Araujo, E. A. d. C., de Andrade, D. F., e Bortolotti, S. L. V. (2009). Teoria da resposta ao item. *Revista da Escola de Enfermagem da USP*, *43*(1), 1000–1008. doi: 10.1590/S0080-62342009000500003 [GS Search]

Baker, F. B., e Kim, S.-H. (2017). *The basics of item response theory using r*. Springer. doi: 10.1007/978-3-319-54205-8 [GS Search]

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

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. [GS Search]

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. doi: 10.1016/j.eswa.2007.03.010 [GS Search]

Chen, C.-M., Lee, H.-M., e Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, *44*(3), 237–255. doi: 10.1016/j.compedu.2004.01.006 [GS Search]

El Falaki, B., El Faddouli, N.-E., Idrissi, M. K., e Bennani, S. (2013). Individualizing hci in e-learning through assessment approach. *The International Journal of Engineering Education*, *29*(3), 650–659. [GS Search]

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. doi: 10.34105/j.kmel.2011.03.025 [GS Search]

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

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

Henderson, M., Ajjawi, R., Boud, D., e Molloy, E. (2019). *The impact of feedback in higher education: Improving assessment outcomes for learners*. Springer Nature. [GS Search]

Iahad, N., Dafoulas, G. A., Kalaitzakis, E., e Macaulay, L. A. (2004). Evaluation of online assessment: The role of feedback in learner-centered e-learning. *37th Annual Hawaii International Conference on System Sciences*. doi: 10.1109/HICSS.2004.1265051 [GS Search]

Isotani, S., e de Oliveira Brandão, L. (2004). Ferramenta de avaliação automática no igeom. *Simpósio Brasileiro de Informática na Educação-SBIE*, 319–328. [GS Search]

Juniwal, G. (2014). *Cpsgrader: Auto-grading and feedback generation for cyber-physical systems education*. EECS department, University of California, Berkeley. [GS Search]

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

Lord, F. (1952). A theory of test scores. *Psychometric monographs*. [GS Search]

Meyer, A. I. d. S., e Mont'Alverne, C. R. d. S. A. (2021). Proposta pedagógica do moodle. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, *7*, 226–241. doi: 10.51891/rease.v7i5.1187 [GS Search]

Moreira, M. P., e Favero, E. L. (2009). Um ambiente para ensino de programação com feedback automático de exercícios. *Workshop sobre Educação em Computação*. [GS Search]

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90–100. doi: 10.1016/S0022-2496(02)00028-7 [GS Search]

Paek, I., e Cole, K. (2019). *Using r for item response theory model applications*. Routledge. [GS Search]

Petrassi, A. C. A., Bornia, A. C., e Andrade, D. F. (2021). Avaliação do nível de satisfação discente de uma instituição de ensino superior: uma análise dos métodos da teoria clássica da medida e da teoria da resposta ao item. *Ensaio: Avaliação e Políticas Públicas em Educação*. doi: 10.1590/S0104-40362021002902192 [GS Search]

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. [GS Search]

Pimentel, E. P., Real, E. M., Braga, J. C., e Botelho, W. T. (2020). Análise dos resultados de insucesso escolar com o suporte de mineração de processos educacionais. *Simpósio Brasileiro de Informática na Educação*, 132–141. doi: 10.5753/cbie.sbie.2020.132 [GS Search]

Rajamani, K., e Kathiravan, V. (2013). An adaptive assessment system to compose serial test sheets using item response theory. *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 120–124. doi: 10.1109/ICPRIME.2013.6496458 [GS Search]

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. [GS Search]

Santo, J. d. E., Castelano, K., e Almeida, J. d. (2012). Uso de tecnologias na prática docente: um estudo de caso no contexto de uma escola pública do interior do Rio de Janeiro. *II Congresso Internacional TIC e Educação. Universidade Tecnológica Federal do Paraná. Espírito Santo: Revista Educação & Tecnologia*, 1023–1031. [GS Search]

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. [GS Search]

Spearman, C. (1961). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. doi: 10.1037/11491-006 [GS Search]

Umrani-Khan, F., e Iyer, S. (2009). ELAM: a model for acceptance and use of e-learning by teachers and students. *Proceedings of the International Conference on e-Learning*, 475–485. [GS Search]

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

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*, 475–511. doi: 10.3102/0034654314564881 [GS Scholar]

Veloso, B. (2021). Paulo Freire e educação a distância: visão propositiva para explorar a autonomia no ensino-aprendizagem. *Congresso Brasileiro de Ensino Superior a Distância*. [GS Search]

Wright, B. D., e Mead, R. J. (1980). Bical: Calibrating items and scales with the rasch model. *Research memorandum*, *23*. [GS Search]

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