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: 13/04/2020; 1st round notif.: 22/05/2020; New version: 04/08/2020; 2nd round notif.: 31/08/2020;
Camera ready: 06/09/2020; Edition review: 16/10/2020; Available online: 17/10/2020; Published: 17/10/2020;

# Estudo Comparativo entre Modelos que Estimam a Habilidade dos Estudantes em Ambientes Virtuais de Programação

## *Title: Comparative Study between Models that Estimate Student's Skill in Virtual Programming Environments*

*Fabiana Zaffalon*
*Universidade Federal do Rio Grande*
*fabinhazaffalon@gmail.com*

*André Prisco*
*Universidade Federal do Rio Grande*
*prisco.c3@gmail.com*

*Ricardo Souza*
*Universidade Federal do Rio Grande*
*rcrdsou@hotmail.com*

*Jean Luca Bez*
*Universidade Regional do Alto Uruguai e das Missões*
*bez@urionlinejudge.com.br*

*Neilor Tonin*
*Universidade Regional do Alto Uruguai e das Missões*
*neilor@urionlinejudge.com.br*

*Rafael Penna*
*Universidade Federal do Rio Grande*
*rapennas@gmail.com*

*Silvia Botelho*
*Universidade Federal do Rio Grande*
*silviacb.botelho@gmail.com*

## Resumo

*É crescente o número de plataformas online que disponibilizam exercícios de programação, onde os estudantes submetem a resolução destes exercícios e recebem um feedback automático do sistema, sem intervenção humana. Esses ambientes permitem o registro de muitos aspectos das submissões e, dessa forma, os modelos de avaliação educacional podem ser utilizados para inferir as habilidades trabalhadas em cada solução. Neste trabalho apresentamos uma análise comparativa de três modelos que estimam a habilidade dos estudantes: Elo, Teoria de Resposta ao Item (TRI) e M-ERS (Multidimensional Extension of the ERS). O Elo foi desenvolvido para classificar jogadores de xadrez, através do histórico de jogo, mas foi adaptado para estimar a habilidade dos estudantes através do histórico de submissões dos problemas. A TRI estima a habilidade através de um conjunto de respostas dadas a um conjunto de itens, existem alguns modelos de TRI que variam de acordo com o tipo de resposta. M-ERS é uma adaptação do Elo e TRI que combina os dois modelos e rastreia as múltiplas habilidades dos estudantes. Os modelos Elo, TRI de 2 parâmetros, TRI de resposta gradual e o M-ERS foram aplicados em uma base de dados disponibilizada por uma plataforma Online Judge. Os resultados obtidos apontam diferenças entre os modelos em relação às habilidades estimadas, diferenças que acredita-se estar relacionadas à forma com que cada modelo estima os parâmetros.*
***Palavras-chave:*** *Habilidade; TRI; Elo; M-ERS.*

## Abstract

*The number of online platforms offering programming exercises is increasing, where students submit exercise resolutions and receive automatic feedback from the system, without human intervention. These environments allow the recording*

*Cite as: Zaffalon, F.; Prisco, A.; Souza, R.; Bez, J. L.; Tonin, N.; Penna, R. & Botelho, S. (2020). Comparative Study between Models that Estimate Student's Skill in Virtual Programming Environments (Estudo Comparativo entre Modelos que Estimam a Habilidade dos Estudantes em Ambientes Virtuais de Programação). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação – RBIE), 28, 776-795. DOI: 10.5753/RBIE.2020.28.0.776.*

*of many aspects of the submissions and, thus, the educational evaluation models can be used to infer the skills worked in each solution. In this paper we present a comparative analysis of three models that estimate student's skill: Elo, Item Response Theory (IRT) and M-ERS (Multidimensional Extension of the ERS). Elo was developed to classify chess players, through their game history, but it was adapted to estimate the student's skill through the history of problem submissions. The IRT estimates the skill through a set of answers given to a set of items, there are some IRT models that vary according to the type of response. M-ERS is an adaptation of Elo and IRT that combines the two models and tracks the multiple skills of students. The Elo models, 2-parameter IRT, gradual response IRT and M-ERS were applied to a database provided by an Online Judge platform. The results obtained point out differences between the models regarding the estimated skills, differences that are believed to be related to the way in which each model estimates the parameters.*
***Keywords:*** *Skill; IRT; Elo; M-ERS.*

# 1 Introdução

Os modelos de avaliação educacional com maior destaque são aqueles que visam apresentar dados precisos sobre a construção das competências, sejam de estudantes presenciais ou a distância através do uso de plataformas *online*. Como muitas dessas plataformas utilizam um sistema de avaliação e $feedback$ automáticos, os educadores buscam a melhoria dos modelos para avaliar a habilidade dos estudantes.

Muitos exercícios práticos exigem mais de uma habilidade para que possam ser resolvidos. Os estudantes da área de computação, além da habilidade da programação em si, referente às particularidades da linguagem de programação, também necessitam da habilidade matemática, do raciocínio lógico, de capacidade de interpretação, entre outras (Moreira et al., 2018) e (Robins, 2010). Para esses estudantes há plataformas que oferecem materiais e exercícios direcionados para o aprendizado de programação.

Perrenoud and Magne (1999) definem competência como o domínio global ou prático de uma situação cotidiana, e habilidade como domínio de uma operação específica ou ações que atendam à uma ou mais competências. Tais definições vem ao encontro de Moretto (França, 2020) que define habilidade no âmbito da educação como "aplicação prática de uma determinada competência para resolver uma situação complexa". No presente trabalho assume-se como sub-habilidade as várias habilidades necessárias para resolver determinado problema.

Diante desse cenário, o propósito deste trabalho é analisar e comparar os dados obtidos a partir de modelos que estimam as habilidades dos estudantes: Elo, Teoria de Resposta ao Item (TRI) e M-ERS. Como critério de análise adotado utiliza-se a comparação dos valores de habilidades dos estudantes gerados pelos modelos.

O modelo Elo é um sistema de classificação estatística criado, originalmente, para classificar jogadores de xadrez através dos seus históricos de jogo (Elo, 1978). Ao invés de relacionar dois jogadores, conforme Elo original, nesse trabalho utiliza-se uma adaptação do Elo proposto por Pelánek (2016), que adequou a métrica para relacionar um estudante ao problema com o qual interage, ou seja, um estudante compete com um problema.

A TRI busca representar a probabilidade de um indivíduo acertar a resposta referente a um item. A relação entre a probabilidade de um indivíduo dar uma resposta correta a um item, se expressa de forma que, quanto maior a habilidade do indivíduo, maior será a probabilidade de acerto (Andrade et al., 2000). Existem vários modelos de TRI que dependem da natureza do item, número de população envolvida e quantidade de habilidades a serem estimadas (Andrade et al., 2000) e (Baker, 2001).

O M-ERS é um modelo que adapta o modelo Elo a um modelo compensatório de TRI Multidimensional, tendo como objetivo avaliar concomitantemente as várias habilidades dos estudantes. A ideia é que ao invés de assumir um traço unidimensional de respostas aos itens, essa abordagem assume que um único item pode envolver mais de uma habilidade (Park et al., 2019).

Para aplicar os modelos, foi utilizado um banco de dados de uma plataforma *OnlineJudge*, que contém problemas de programação onde os usuários (aprendizes) resolvem os problemas e recebem um $feedback$ de acerto ou erro. As aplicações dos modelos têm por finalidade entender as relações entre eles.

De acordo com os resultados gerados por cada modelo, a análise e comparação foram feitas em dois grupos: 1) os modelos da TRI (TRI Logístico de 2 Parâmetros com a TRI resposta gradual), por retornarem um valor escalar para cada habilidade; e 2) Elo com o M-ERS, por retornarem o histórico das habilidades de cada estudante.

Nas seções seguintes são apresentados o referencial teórico que embasou a pesquisa, a metodologia utilizada para a execução dos experimentos, a apresentação e discussão dos resultados obtidos e, por fim, as conclusões.

## 2 Referencial Teórico

### 2.1 Sistema de Classificação Elo

Originalmente, o Elo[1] classifica jogadores de xadrez atribuindo a cada jogador um *score* inicial (um valor inicial de Elo) e, à medida que vai participando dos jogos, esse Elo vai sendo atualizado de acordo com os resultados. O modelo trabalha em função da probabilidade esperada e do resultado: caso o resultado atenda à probabilidade de que um jogador com maior habilidade vença o jogador com menor, os valores de Elo recebem pequenas atualizações; caso contrário, a atualização é maior (Elo, 1978) e (Pelánek, 2016).

A probabilidade de que o jogador $i$ ganhe do jogador $j$ $(R_{ij} = 1)$ é apresentada na Equação 1 (Pelánek, 2016).

$$P(R_{ij} = 1) = \frac{1}{1 + 10^{\frac{\theta_j - \theta_i}{400}}} \tag{1}$$

onde $R = \{0, 1\}$ é o conjunto de resultados de um jogo: 1 (ganhar) e 0 (perder), dado um jogo entre os jogadores $i$ e $j$, com um Elo $\theta_i$ e $\theta_j$, respectivamente. Ao final da partida novos Elos são calculados de acordo com as probabilidades esperadas dos resultados e de acordo com os Elos anteriores. A constante $k$ indica a escala de atualização do valor Elo, quanto maior o $k$, maior será a variação, conforme Equação 2 (Pelánek, 2016).

$$\theta_i = \theta_i + k(R_{ij} - P(R_{ij} = 1)) \tag{2}$$

O Elo vem sendo utilizado como método de avaliação em ambientes educacionais, onde entende-se haver uma relação entre o estudante e o problema que ele está resolvendo, estimando,

[1] sobrenome do seu criador: Arpad Emmerich Elo.

assim, a habilidade do estudante e a dificuldade do problema (Pelánek, 2016). A estimativa acontece de maneira contínua, pois a atualização dos *scores* acontece ao término de cada evento.

Em Prisco et al. (2018), os autores apresentam uma adaptação do Elo semelhante à usada na escolha de oponentes em torneios de xadrez ou partidas *online*. Essa abordagem, em vez de relacionar um jogador a outro jogador, associa um estudante a um problema. Assim, cada submissão de problema é considerada um jogo, ou seja, um "duelo" entre o estudante e o problema. Em cada início de submissão, o algoritmo armazena o Elo dos dois e, após a avaliação do "duelo", os Elos, tanto do estudante quanto do problema, são atualizados.

### 2.2 Teoria de Resposta ao Item (TRI)

A Teoria de Resposta ao Item (TRI) é considerada um importante recurso no processo quantitativo de avaliação educacional. Permite uma análise mais precisa de cada item que compõe o instrumento de avaliação, levando em consideração as características dos itens na produção das habilidades (Andrade et al., 2000) e (Baker, 2001).

A TRI compreende um conjunto de modelos matemáticos construídos para representar a probabilidade de um estudante responder corretamente um item em determinado teste. Os modelos estimam as habilidades $\theta$ dos estudantes através das respostas dadas ao conjunto de itens em determinada área de conhecimento a ser avaliada. A estimativa da habilidade é relacionada com a probabilidade do estudante acertar o item/problema, considerando um ou mais parâmetros. Dessa forma, quanto maior a habilidade do estudante, maior será a probabilidade de ele acertar o problema (Andrade et al., 2000).

Basicamente, os modelos da TRI, dependem de três fatores: natureza do item (dicotômicos ou não dicotômicos), número de populações envolvidas (uma ou mais) e quantidade de traços latentes (habilidades) a ser medida (uma ou mais) (Andrade et al., 2000) e (Baker, 2001). Para itens dicotômicos, a literatura apresenta três modelos que diferem entre si pela quantidade de parâmetros utilizados para descrever o item (Andrade et al., 2000) e (Baker, 2001):

- Modelo Logístico de 1 parâmetro (modelo de Rasch): a dificuldade do item.
- Modelo Logístico de 2 parâmetros (ML2): a dificuldade e a discriminação do item.
- Modelo Logístico de 3 parâmetros (ML3): a dificuldade e a discriminação do item, e a probabilidade de acerto ao acaso.

O ML3, Equação 3, calcula a probabilidade do indivíduo com habilidade $\theta$ de acertar o item $j$, levando em consideração a discriminação do item $a$, dificuldade do item $b$ e a chance de acerto ao acaso $c$ (Baker, 2001).

$$P(\theta) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta - b_j)}} \tag{3}$$

O parâmetro de discriminação do item $a_j$ indica o quanto um item distingue os indivíduos com diferentes níveis de habilidade, supondo que um mesmo teste será aplicado a diferentes indivíduos com diferentes habilidades. Assim, esse parâmetro pode aumentar, ou não, a diferença entre as probabilidades de estudantes com habilidades distintas responderem corretamente o item. Itens com maiores valores de $a_j$ fornecem melhores discriminações (Tavares, 2014).

A dificuldade do item $b_j$ é expresso na mesma escala da habilidade, ou seja, representa a habilidade que o indivíduo precisa possuir para responder corretamente o item (Oliveira, 2017) e (Araujo et al., 2009).

O parâmetro de acerto ao acaso $c_j$ é a probabilidade de um indivíduo com baixa habilidade acertar casualmente um item. Dessa forma, caso não seja permitido responder ao acaso, como por exemplo exercícios de programação, o parâmetro $c_j$ assume valor 0 (zero) (Andrade et al., 2000) e (Oliveira, 2017).

Na TRI, a habilidade tem relação com a probabilidade do indivíduo responder corretamente um item. A Curva Característica do Item (CCI), Figura 1, representa a relação existente entre a habilidade do indivíduo e o seu desempenho nos itens: indivíduos com maiores valores de habilidades possuem maior probabilidade de responder corretamente ao item (Andrade et al., 2000), (Andrade & Justino, 2007) e (Baker, 2001).

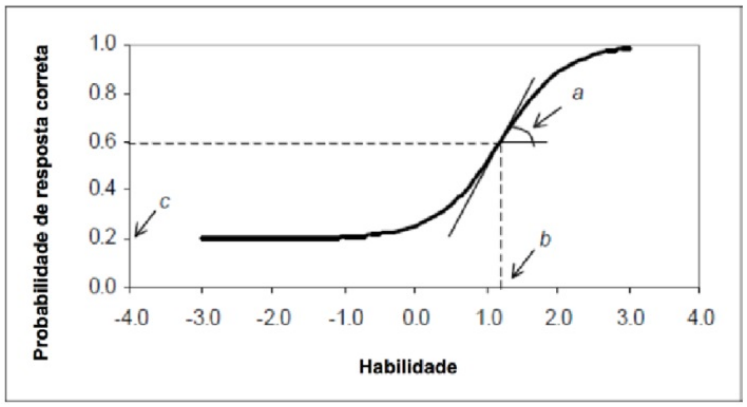



Figura 1: Curva Característica do Item (Andrade et al., 2000).

Os modelos para itens não dicotômicos são aplicados a testes com itens abertos (resposta livre) ou com itens de múltipla escolha (avaliados de forma graduada). Um dos modelos para esse tipo de item é o de resposta gradual, de Samejima, que é uma generalização da TRI ML2 e assume que as categorias de respostas podem ser ordenadas entre si, como uma escala de Likert (Braga, 2015).

Neste modelo, supõe-se que os *scores* das categorias de um item $i$ estejam dispostos em ordem crescente e indicados por $k = 0,1,...,m_i$, onde $(m_i + 1)$ é o número de categorias do $i$-ésimo item (Braga, 2015). A probabilidade de um indivíduo $j$ escolher uma determinada categoria do item $i$ é representada pela Equação 4 (Braga, 2015) e (Soares et al., 2004).

$$P_{i,k}(\theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_{i,k})}} - \frac{1}{1+e^{-Da_i(\theta_j - b_{i,k+1})}} \tag{4}$$

onde $i = 1,...,I$, $j = 1,...,n$ e $b_{i,k}$ é o parâmetro de dificuldade da $k$-ésima categoria do item $i$. Demais modelos para itens não dicotômicos podem ser encontrados em (Andrade et al., 2000) e (Braga, 2015).

Tais modelos consideram que o teste seja um instrumento unidimensional que implica na existência ou predominância de apenas uma habilidade, o que não se aplica em muitas situações práticas. Um teste de matemática, por exemplo, pode exigir a interpretação de texto antes mesmo de exigir o desenvolvimento matemático e, nesse caso, trata-se de um teste bidimensional, pois requer duas habilidades (Nojosa, 2002).

Pesquisas vêm indicando que o modelo TRI Multidimensional (TRIM) se adapta melhor aos dados reais do que os modelos unidimensionais, pois na educação as respostas dos sujeitos são determinadas por mais de uma habilidade (Pasquali, 2018).

Os modelos TRIM são divididos em duas classes: compensatórios e não compensatórios. Um modelo é considerado compensatório quando a probabilidade de acertar o item é mantida ou aumentada, mesmo que uma das habilidades seja baixa, sendo essa compensada por outra habilidade mais alta (Nojosa, 2002). O modelo compensatório da TRIM é apresentado pela Equação 5 (Reckase, 2006).

$$P(u_i = 1|\theta_j) = \frac{e^{a_{ik}\theta_{jk}+d_i}}{1+e^{a_{ik}\theta_{jk}+d_{i'}}} \tag{5}$$

onde $u_i$ é a resposta ao item $i$; $a_{ik}$ é o parâmetro de discriminação do item $i$ na dimensão $k$; $\theta_{jk}$ é o traço latente (habilidade) do indivíduo $j$ na dimensão $k$ e $d_i$ é um escalar que indica a dificuldade do item $i$. O expoente $e$ na Equação 5 pode ser escrito de acordo com a Equação 6 (Reckase, 2006).

$$a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + ... + a_{im}\theta_{jm} + d_i \tag{6}$$

Independente do modelo, um dos principais passos da TRI é a estimação dos parâmetros dos itens (calibração) e das habilidades, que podem ser feitas utilizando os Métodos Bayesiano ou o de Máxima Verossimilhança Marginal (Andrade et al., 2000), (Araujo et al., 2009) e (Chalmers, 2012).

O Método Bayesiano estabelece distribuições a priori para os parâmetros de interesse, constrói uma nova função denominada distribuição a posteriori e estima os parâmetros com base

em alguma característica dessa distribuição. Já o Método de Máxima Verossimilhança Marginal, através de iterações, estima os parâmetros (dos itens e habilidade) em duas etapas: na primeira, o processo de estimação dos parâmetros começa a partir das respostas dos indivíduos, assumindo uma certa distribuição para as habilidades; na segunda, assumindo que os parâmetros dos itens já são conhecidos, estima as habilidades. Essa iteração ocorre até que já não haja mais variações significativas nas estimativas (Andrade et al., 2000), (Araujo et al., 2009) e (Chalmers, 2012).

Para Pelánek (2016), existe uma relação entre o sistema de classificação Elo e a TRI modelo de Rasch. O que os difere é o procedimento de estimativa dos parâmetros e em suas suposições básicas: TRI assume que a habilidade do estudante é constante, enquanto o Elo foi implementado para rastrear as mudanças nos níveis de habilidades.

### 2.3 Modelo M-ERS

M-ERS (*Multidimensional Extension of the ERS*) é um modelo proposto por Park et al. (2019), para aplicar em sistemas adaptativos de aprendizagem, que incorpora o modelo compensatório da TRIM ao modelo Elo.

M-ERS atualiza várias habilidades dos estudantes, de forma simultânea, baseado em um modelo compensatório da TRIM, ao contrário do que ocorre no modelo de Rasch, que considera a existência de apenas uma habilidade para a resolução dos itens. A probabilidade $P_{ij}$ do estudante $i$ acertar o item $j$ da TRIM compensatória é utilizada no modelo Elo para atualizar as habilidades dos estudantes e a dificuldade dos itens.

Em um modelo conjuntivo pressupõe-se que o estudante deva ter cada uma das habilidades relevantes para que ele possa responder corretamente um determinado item. Já, em um modelo compensatório, supõe-se que a falta de uma habilidade pode ser compensada por outra habilidade de maior nível, conforme Equação 7 (Park et al., 2019).

$$P_{ij} = P(Y_{ij} = 1) = \frac{exp(\sum_{m=1}^{M} \alpha_{jm}\theta_{im} - \beta_j)}{1 + exp(\sum_{m=1}^{M} \alpha_{jm}\theta_{im} - \beta_j)'} \tag{7}$$

onde $\theta_{im}$ é a habilidade $m$ do estudante $i$ ($m$=*1, .., M*), sendo $M$ o número de habilidades, $\alpha_{jm}$ é a discriminação do item $j$ correspondente à $m$ dimensão de habilidade e $\beta_j$ é o nível de dificuldade do item $j$.

A diferença entre o desempenho observado $Y_{ij}$ e o desempenho esperado $P_{ij}$, com base nos modelos da TRIM, é usada para atualizar as habilidades após cada resposta do item. $P_{ij}$, dentro do Elo, para a $m$-ésima habilidade do estudante $i$ no tempo $t$ é atualizado conforme Equação 8 (Park et al., 2019).

$$\begin{aligned} \hat{\theta}_{im(t)} &= \hat{\theta}_{im(t-1)} + D_{m(t)}K\{Y_{ij(t)} - P_{ij(t)}\} \\ \hat{\beta}_{j(t)} &= \hat{\beta}_{j(t-1)} - D_{m(t)}K\{Y_{ij(t)} - P_{ij(t)}\} \end{aligned} \tag{8}$$

onde *t* é a medição atual, *t-1* é a medição anterior, $D_{m(t)}$ é um peso para especificar se a habilidade *m* é indicada pelo item dado no *t*-ésimo tempo. Para a habilidade que é indicada pelo item, $D_{m(t)}$ é igual a 1. Para a habilidade que não é indicada pelo item, o peso leva valores entre 0 e 1. *K* diminui linearmente, entre 0,4 e 0,1, em função do número total de itens respondidos. Uma vez conhecido o valor da discriminação do item esse não é atualizado no modelo Elo.

### 2.4 *Online Judges*

*Online Judges* são sistemas voltados para avaliação de códigos-fonte de algoritmos enviados pelos usuários ou estudantes (Wasik et al., 2018). A plataforma URI *Online Judge* (Bez et al., 2014) contém uma base de problemas de programação classificadas por níveis. Muitos professores utilizam o URI *Online Judge* como uma ferramenta didática, criando suas salas virtuais e selecionando os exercícios disponibilizados pela plataforma. Para resolver tais problemas, os estudantes devem elaborar um algoritmo (programa) em uma das linguagens de programação aceitas pela plataforma e submetê-lo. O gráfico da Figura 2a ilustra as linguagens mais utilizadas entre os usuários da plataforma URI (Bez et al., 2011).

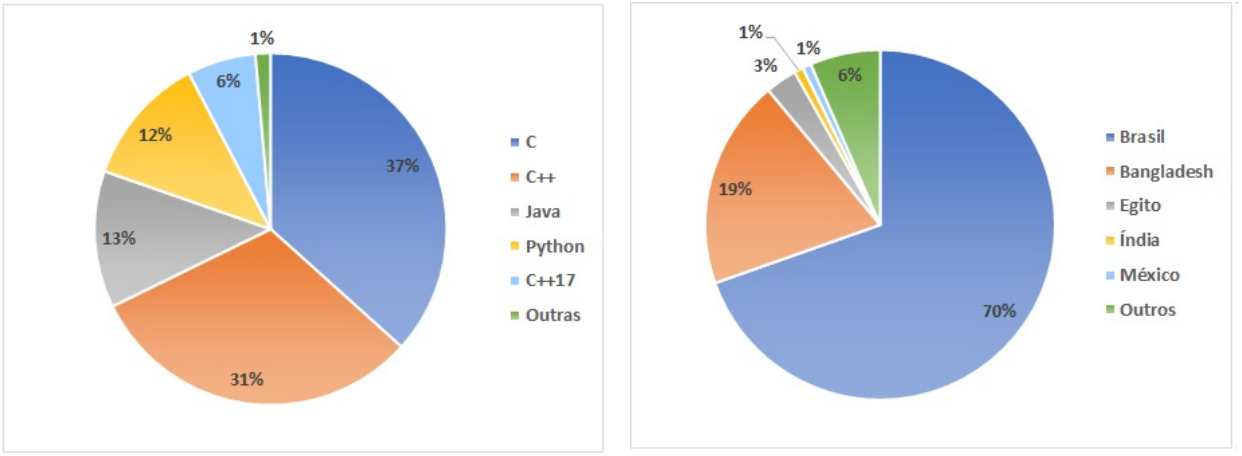
A plataforma URI tem um sistema automático de avaliação que faz a análise do código-fonte submetido e dá um $feedback$ ao estudante. Os tipos de $feedback$ são: aceito (problema aceito sem erro) ou erro (compilação, execução, apresentação, tempo de execução, tempo excedido, falha de comunicação com o servidor ou resposta errada).

Para que os usuários possam escolher exercícios específicos, os problemas são classificados em: Iniciante, *Ad-Hoc*, *Strings*, Estruturas de Dados, Paradigmas, Matemática, Gráfico e Geometria Computacional (Bez et al., 2014). O usuário pode submeter uma quantidade indeterminada de resolução para o mesmo problema, mesmo a resposta estando correta ou, até acertá-lo.

Atualmente, há estudantes de 241 países cadastrados na plataforma, todos vinculados à instituições de ensino. São 1992 instituições cadastradas e 484 têm mais de 1.000 problemas resolvidos. Os 5 países com maior número de exercícios resolvidos estão ilustrados no Gráfico da Figura 2b (Bez et al., 2011).

## 3 Metodologia

Nesta seção será apresentada a metodologia adotada para os experimentos que foram feitos em uma base de dados de exercícios de programação disponibilizada pelo URI *Online Judge*. Nessa base, foram aplicados os modelos TRI ML2 com a TRI resposta gradual, e modelo Elo com o modelo M-ERS, com o objetivo de analisar e comparar os resultados obtidos através desses modelos, acerca das habilidades dos estudantes.




(a) Linguagens mais utilizadas. (b) Países com mais estudantes cadastrados.

Figura 2: Informações sobre a plataforma URI *OnlineJudge*.

### 3.1 Base de dados

A base de dados foi disponibilizada pela plataforma URI *Online Judge*, composta por 1.048.575 submissões de 62.976 usuários a 1.162 problemas de programação, classificados no nível Iniciante. Os dados disponibilizados foram: id do usuário, id do problema, resposta, data e hora da submissão.

Foram excluídos da base: usuários que acertaram ou erraram todos os problemas submetidos, usuários que submeteram menos de 50 soluções e problemas que tinham menos de 100 submissões. Assim, a base ficou constituída por 210.178 submissões de 405 problemas submetidos por 2.100 usuários.

### 3.2 Tratamento da base de dados

Os identificadores dos usuários foram anonimizados. Os dados textuais recebidos da plataforma URI *Online Judge*, foram transformados em dados numéricos: as respostas corretas receberam valor 1 (um) e os problemas que não foram aceitos receberam valor 0 (zero).

Para aplicar os modelos Elo e M-ERS os dados foram organizados por ordem cronológica de submissão e para os modelos da TRI os dados foram tabulados. Para aplicar a TRI ML2, as informações foram tabuladas de forma a relacionar o usuário aos problemas e às respectivas respostas. Mesmo o usuário tendo submetido várias soluções ao mesmo problema, se em alguma dessas submissões a solução estivesse correta, o problema foi considerado correto. A Figura 3 ilustra um exemplo dos dados tabulados, em que as linhas correspondem aos IDs dos usuários e as colunas aos IDs dos problemas.

Para o modelo da TRI resposta gradual, foram criadas escalas de submissões para cada usuário baseado na escala Likert. A Figura 4 apresenta um exemplo de tabulação da base de

| IDs dos usuários | IDs dos problemas | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1001 | 1002 | 1003 | 1004 | 1005 | 1006 | 1007 | 1008 | 1009 | 1010 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 4 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 5 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |

Figura 3: Exemplo dos dados tabulados para TRI ML2.

dados em que os usuários foram relacionados aos problemas e às respectivas escalas. As linhas correspondem aos IDs dos usuários e as colunas aos IDs dos problemas.

| IDs dos usuários | IDs dos problemas | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1001 | 1002 | 1003 | 1004 | 1005 | 1006 | 1007 | 1008 | 1009 | 1010 |
| 1 | 1 | 5 | 1 | 1 | 5 | 2 | 1 | 2 | 1 | 1 |
| 2 | 1 | 2 | 1 | 3 | 5 | 5 | 4 | 1 | 2 | 2 |
| 3 | 1 | 3 | 2 | 1 | 5 | 4 | 1 | 1 | 2 | 5 |
| 4 | 1 | 2 | 1 | 1 | 2 | 5 | 1 | 2 | 3 | 4 |
| 5 | 1 | 5 | 1 | 4 | 5 | 5 | 2 | 3 | 4 | 1 |

Figura 4: Exemplo dos dados tabulados para TRI resposta gradual.

As escalas abaixo foram criadas de acordo com o número de submissões. Entre as escalas 2 e 5, em alguma submissão a solução estava correta.

- Escala 1 - submeteu, uma ou mais vezes, mas não obteve acerto. Com 33.875 ocorrências.
- Escala 2 - mais de dez submissões. Com 143 ocorrências.
- Escala 3 - de seis a dez submissões. Com 1.063 ocorrências.
- Escala 4 - de duas a cinco submissões. Com 20.545 ocorrências.
- Escala 5 - uma submissão. Com 94.750 ocorrências.

## 3.3 Experimentos

Foram realizados quatro experimentos na mesma base de dados: TRI ML2, Modelo TRI resposta gradual, Modelo Elo e Modelo M-ERS.

### *3.3.1 Modelo TRI ML2*

Foi utilizado o *software* RStudio e o pacote *mirt* que estima os parâmetros dos itens e dos indivíduos, utilizando o método Máxima Verossimilhança Marginal (Chalmers, 2012).

Através do pacote *mirt*, foi feita a calibração e a parametrização para obter o nível de dificuldade e a discriminação dos itens. Após, foi realizada uma revisão dos valores considerados críticos de acordo com as características da TRI. Tal análise proporciona a validação dos itens considerados satisfatórios, bem como a exclusão ou reavaliação dos itens que não se adéquam para a aplicação do modelo (Araujo et al., 2009). Todos os itens analisados foram satisfatórios. Após a estimativa dos parâmetros dos itens, o próximo passo foi estimar as habilidades dos usuários, também através do pacote *mirt*.

### *3.3.2 Modelo TRI resposta gradual*

Com o uso do pacote *mirt*, no *software* RStudio, foi feita a calibração e a parametrização dos itens. Foi efetuada a mesma validação dos valores realizada na subseção 3.3.1, em que todos os valores foram considerados satisfatórios. De posse dos parâmetros dos itens, as habilidades dos usuários foram estimadas através do pacote *mirt*.

### *3.3.3 Modelo Elo*

Os valores Elo de cada usuário foram inicializados com o valor 0 (zero). Para cada problema foi atribuído o valor de dificuldade estimado na TRI ML2. O algoritmo simula cada submissão realizada em ordem cronológica através da resposta de cada envio. Após as estimativas, os novos valores da habilidade Elo dos usuários e da dificuldade dos itens são armazenados no banco de dados. Dessa forma, tem-se o histórico do Elo de cada usuário e dificuldade de cada problema.

### *3.3.4 Modelo M-ERS*

Os parâmetros dos problemas, discriminação e dificuldade, foram estimados através da TRI ML2 e, para a habilidade dos usuários, inicialmente foi atribuído o valor 0 (zero). Como esse modelo leva em consideração as múltiplas habilidades (sub-habilidades) necessárias para resolução dos problemas, cada sub-habilidade dos usuários recebeu o valor 0 (zero) e cada problema foi analisado por três especialistas que identificaram e classificaram essas sub-habilidades (Prisco et al., 2018).

Cada sub-habilidade dos problemas recebeu uma relevância, entre 0 (zero) e 1(um), que indica o quão ela é importante para a interação com o problema. Independentemente do nível de dificuldade, a relevância igual a 1 (um) indica que a sub-habilidade é crítica para uma resposta correta, enquanto que a relevância igual a 0 (zero) indica que a sub-habilidade não é necessária (Prisco et al., 2018). Valores intermediários indicam que o uso de uma habilidade particular ou conceito é positivo, porém opcional, como também indica que pode ser usado, mas não é o desafio central para a resolução do problema (Prisco et al., 2018).

Foram identificadas as seguintes sub-habilidades nos problemas (Prisco et al., 2018):

- Básico: envolve problemas sequenciais, mais simples, com operadores, condicionais e *loops*. Padrões, abstração e conhecimento da língua estão relacionados a esse conceito.

- Matemática: domínio de conhecimento em álgebra e geometria e como aplicar tais conceitos na resolução dos problemas.
- Modularização: capacidade de dividir um problema em subproblemas para habilitar ou facilitar sua solução.
- Linear: envolve conceitos de pilhas, filas e listas. Em alguns problemas, a relevância é alta, já que o objetivo principal do problema é testar o domínio do estudante em alguma dessas estruturas.
- Não linear: domínio e aplicação de grafos e árvores.
- Avançado: abrange conceitos que vão além lógica de programação básica, é necessário saber técnicas de programação avançadas para a resolução do problema.
- *String*: processamento de dados textuais. Embora possa ser agrupado em estruturas lineares, há uma quantidade de problemas que envolvem características específicas de processamento de texto, *substring*, pesquisa e outros que especificamente envolvem essa estrutura.

De posse dos parâmetros dos problemas (discriminação, dificuldade e relevância das sub-habilidades) foi aplicado o algoritmo do modelo M-ERS. Da mesma forma que o modelo Elo, o algoritmo é executado nos dados em ordem cronológica de submissão. De acordo com as respostas, faz as estimativas e armazena os novos valores no banco de dados.

Os resultados obtidos e as discussões acerca dos experimentos são apresentados na próxima seção.

## 4 Resultados e Discussões

Os modelos TRI ML2 e TRI resposta gradual, como resultado, geraram um *score* único para cada usuário, enquanto que os modelos ELO e M-ERS geraram o histórico das habilidades. Devido a essa diferença no tipo de resultados, foram feitos dois comparativos: TRI ML2 com o Modelo TRI resposta gradual; e o Modelo Elo com o Modelo M-ERS.

### 4.1 TRI ML2 e TRI resposta gradual

Os modelos da TRI mostram como resultado um valor escalar que representa a habilidade $\theta$ de cada usuário. Nos dois modelos da TRI cada usuário recebeu um valor exclusivo de habilidade. Na TRI ML2, as habilidades variaram entre $\cong -2,57$ e $\cong 3,52$. Dos 2.100 usuários, 898 ($\cong 42,7\%$) tiveram aumento na habilidade e 1.202 ($\cong 57,2\%$) tiveram queda.

Na TRI modelo resposta gradual, as habilidades variaram entre $\cong -3,13$ e $\cong 3,51$. Dos 2.100 usuários, 968 ($\cong 46\%$) apresentaram aumento de habilidade e 1.132 ($\cong 53,9\%$) apresentaram queda.

Nos dois modelos, os mesmos usuários apresentaram o maior e o menor valor de habilidade. A Tabela 1 relaciona esses usuários com as quantidades de submissões realizadas (coluna Submissões), quantidade de problemas que cada um submeteu (coluna Problemas), quantidade de soluções corretas (coluna Acertos) e incorretas (coluna Erros), bem como os valores de habilidade nos dois modelos (colunas $\theta$ ML2 e $\theta$ resposta gradual).

Tabela 1: Usuários com maior e menor valores de habilidade nos modelos TRI ML2 e resposta gradual.

| **Usuário** | **Submissões** | **Problemas** | **Acertos** | **Erros** | **$\theta$ ML2** | **$\theta$ resposta gradual** |
|---|---|---|---|---|---|---|
| 1 | 369 | 341 | 340 | 29 | $\cong 3,52$ | $\cong 3,51$ |
| 2 | 84 | 58 | 17 | 67 | $\cong -2,57$ | $\cong -3,13$ |

No modelo TRI resposta gradual, o usuário 1 teve 321 ocorrências na escala 5, ou seja, dos 341 problemas resolvidos, 321 ele obteve acerto na primeira submissão. Na categoria 4 (de 2 a 5 submissões), 19 problemas e somente 1 problema na escala 1 (submeteu 1 ou mais vezes e não acertou). O usuário 2 teve 13 ocorrências na escala 5, 4 ocorrências na escala 4 e 41 ocorrências na escala 1.

A Tabela 2 relaciona alguns usuários escolhidos aleatoriamente para a verificar a diferença na estimativa de valores das habilidades. É possível observar que essas diferenças não são consideradas significativas.

Tabela 2: Relação das habilidades nos modelos TRI ML2 e resposta gradual.

| **Usuário** | **$\theta$ ML2** | **$\theta$ resposta gradual** | **Diferença** |
|---|---|---|---|
| 1 | $\cong 3,52$ | $\cong 3,51$ | $\cong 0,01$ |
| 2 | $\cong -2,57$ | $\cong -3,13$ | $\cong 0,56$ |
| 3 | $\cong 2,99$ | $\cong 3,36$ | $\cong 0,37$ |
| 4 | $\cong 2,51$ | $\cong 2,98$ | $\cong 0,47$ |

Os percentis da diferença das habilidades foram calculados com o objetivo de analisar a distribuição dos usuários de acordo essa diferença, ilustrada na Figura 5. Observa-se que, dos 2.100 usuários, 1.165 (97%) estão na faixa de 10% de diferença nas habilidades estimadas pelo TRI ML2 e resposta gradual; 561 usuários estão na faixa dos 20% de diferença e, assim, sucessivamente. À medida que o percentual de diferença vai aumentando, a quantidade de usuários vai diminuindo, o que demonstra que não há diferença significativa na estimativa das habilidades entre os modelos.

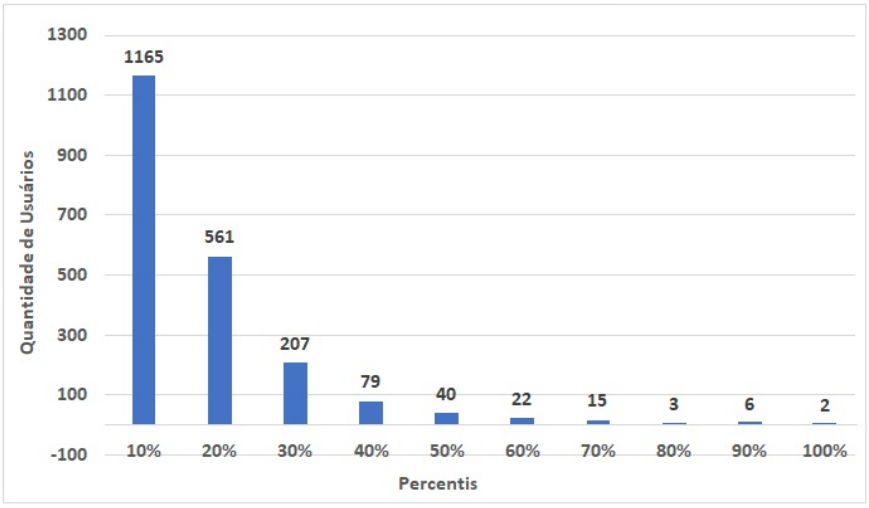



Figura 5: Percentis da diferença entre as habilidades calculados pelos modelos TRI ML2 e TRI resposta gradual.

A Figura 6 apresenta o diagrama de dispersão entre as habilidades estimadas nos dois modelos. O coeficiente de correlação de Pearson aplicado a esses dados foi igual a 0,8505277, o que sugere uma forte correlação.

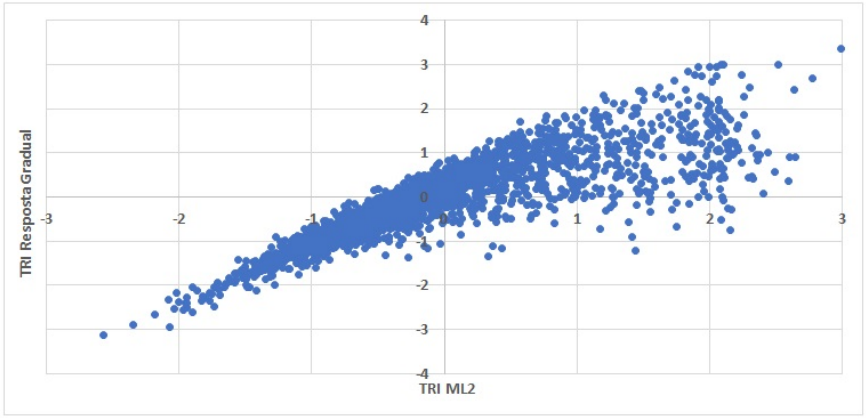



Figura 6: Gráfico de dispersão das habilidades nos modelos TRI ML2 e resposta gradual.

## 4.2 Modelo Elo e M-ERS

Os modelos Elo e M-ERS, ao contrário dos modelos da TRI, apresentam o histórico de habilidades dos usuários, bem como a variação no valor de dificuldade dos problemas. Isso acontece porque os modelos consideram todas as submissões. Quando o usuário envia uma solução correta do problema as habilidades desse usuário são atualizadas positivamente e, por consequência, a dificuldade do problema também é atualizada, porém, negativamente, diminuindo seu valor. O contrário também

acontece, submissão errada eleva a dificuldade do problema e diminui a habilidade do usuário.

No modelo Elo as habilidades variaram entre $\cong -32,27$ e $\cong 53,93$ e no modelo M-ERS variaram entre $\cong -37,72$ e $\cong 13,00$. O coeficiente de correlação de Pearson aplicado aos dados, referentes aos históricos das habilidades Elo e M-ERS, foi igual 0,7695701, o que sugere uma forte correlação. A Figura 7 apresenta o diagrama de dispersão das habilidades Elo e M-ERS.

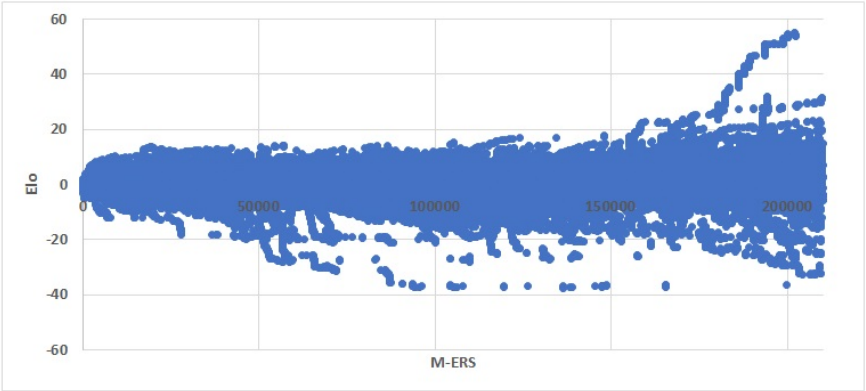



Figura 7: Gráfico de dispersão das habilidades nos modelos Elo e M-ERS.

Foram analisados os históricos das habilidades Elo e M-ERS do usuário 1, conforme ilustra a Figura 8, onde o eixo horizontal apresenta a quantidade de submissões e o eixo vertical apresenta o valor da habilidade. Assim como nos modelos TRI, esse usuário ficou com maior habilidade Elo e M-ERS.

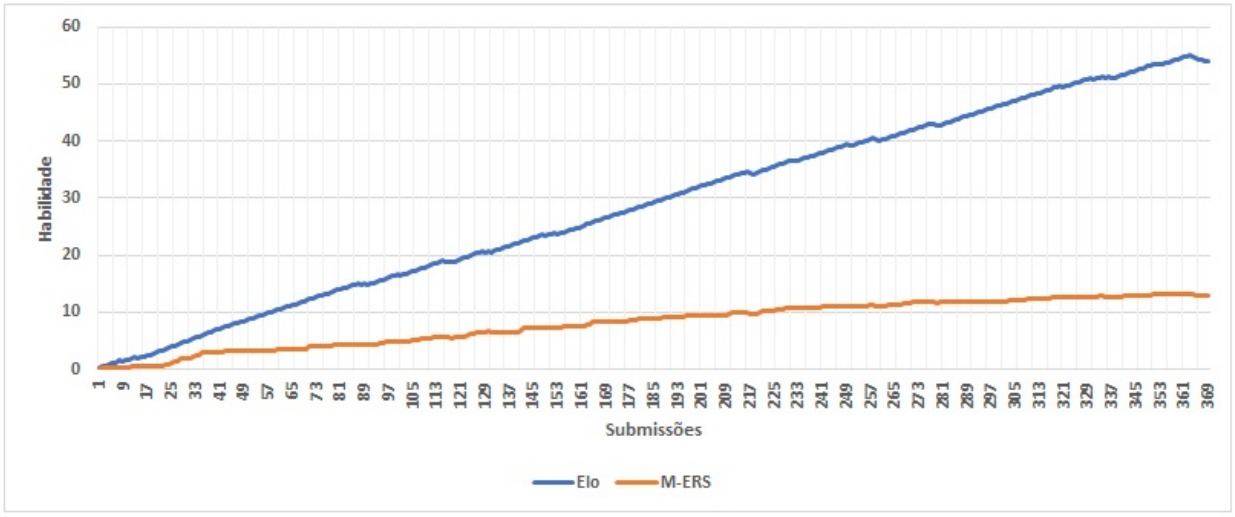



Figura 8: Histórico da habilidade do usuário 1 nos modelos Elo e M-ERS.

Observa-se o aumento no valor das habilidades em ambos os modelos. Apesar das habilidades serem inicializadas em 0 (zero) nos dois modelos, os valores de Elo ficaram superiores. Isso se

dá porque o modelo Elo utiliza a constante ($k$=0,4) para atualizar as habilidades, enquanto que no modelo M-ERS esse valor vai decrescendo linearmente a cada submissão.

O modelo M-ERS também permite analisar o histórico das sub-habilidades dos usuários, conforme Figura 9, onde o eixo horizontal apresenta a quantidade de submissões e o eixo vertical apresenta o valor da habilidade. É possível identificar que inicialmente o usuário 1 resolveu os problemas mais básicos e, partir da 26ª submissão, os problemas resolvidos envolviam as sub-habilidades *String* e Linear.

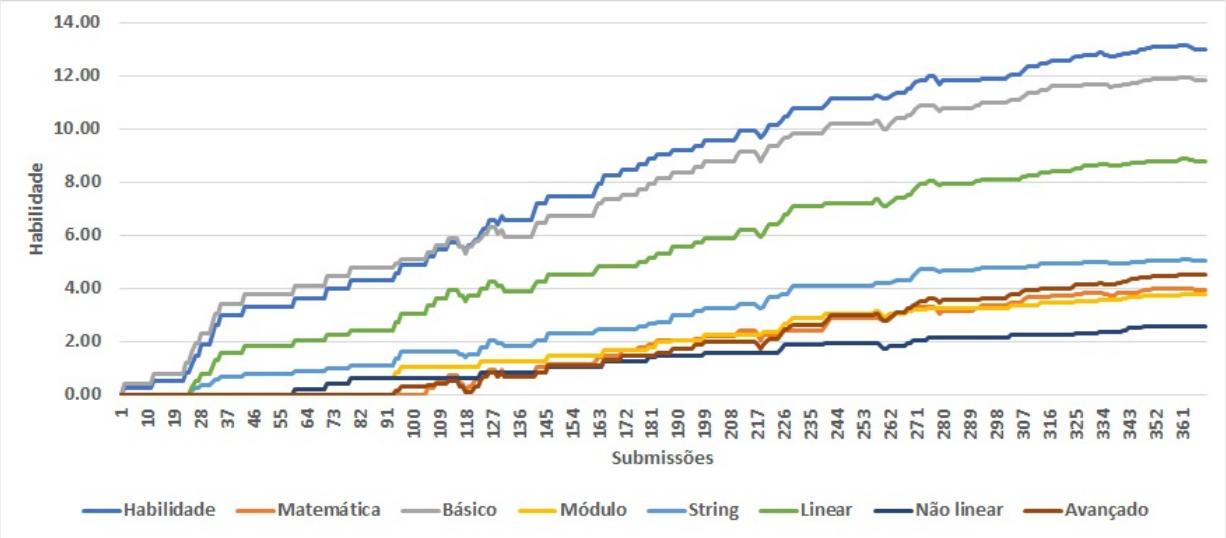



Figura 9: Histórico da habilidade e das sub-habilidades do usuário 1 no modelo M-ERS.

De acordo com o gráfico, da mesma forma que ocorre com o histórico das habilidades no modelo Elo, há uma oscilação no histórico da habilidade do usuário que varia conforme as submissões corretas e incorretas.

Para entender melhor o histórico das sub-habilidades, é necessário observar os valores de relevância nos problemas. A sub-habilidade Básico possui grande relevância em todos os problemas que o usuário submeteu. Por isso, no gráfico, essa linha acompanha a linha da habilidade.

## 5 Conclusões

Para esse artigo, foi feito um comparativo entre quatro modelos que estimam as habilidades de estudantes. De acordo com os resultados obtidos, foram comparados os modelos TRI ML2 com a TRI resposta gradual, por retornarem um valor escalar para cada habilidade. Os modelos Elo e M-ERS foram comparados por retornarem o histórico das habilidades de cada estudante.

Os modelos da TRI baseiam-se na suposição de que uma habilidade é constante ou fixa, não

levam em consideração o desempenho dos usuários ao longo do tempo. Para que se pudesse aplicar a TRI ML2 em uma base de dados em que os usuários podem submeter várias soluções para o mesmo problema, foi preciso tratar os dados de forma a relacionar os usuários com os problemas e a uma reposta apenas. Assim, para cada usuário, analisou-se o conjunto de respostas que ele deu ao mesmo problema: no caso de alguma resposta correta a questão foi considerada correta, independente do número de tentativas até o acerto, caso contrário, foi considerada incorreta.

Para aplicar a TRI resposta gradual, os dados foram organizados baseados na escala Likert, de acordo com o número de submissões até o acerto. Ambos os modelos resultaram um valor escalar para a habilidade de cada usuário. Apesar de serem aplicados na mesma base e tabulados de forma diferente, os resultados obtidos foram próximos e constatou-se uma forte correlação entre eles.

O modelo de classificação Elo e o modelo M-ERS permitem o acompanhamento do histórico das habilidades dos usuários considerando todas interações entre eles e os problemas. Apesar de apresentar uma discrepância em relação aos resultados das habilidades entre os dois modelos, foi constatado que há uma correlação entre eles. Nos dois modelos os usuários perderam habilidade quando erraram o problema e ganharam quando contrário. Embora o M-ERS utilize o modelo Elo para a atualização da habilidade, essa diferença nos resultados se dá devido à forma com que cada modelo atualiza as habilidades dos usuários, o Elo utiliza uma constante ($k$=0,4) enquanto o M-ERS diminui linearmente esse valor a cada submissão.

O diferencial do modelo M-ERS sobre o modelo Elo, se dá na estimativa e no histórico da habilidade e das sub-habilidades dos usuários, uma vez que os exercícios de programação geralmente envolvem mais de uma sub-habilidade. Com o histórico das sub-habilidades é possível identificar em qual delas os usuários apresentam mais dificuldade e qual delas precisam ser mais exercitadas. Esse tipo de resultado permite que haja um acompanhamento mais preciso em relação à evolução dos estudantes.

Os modelos da TRI são aplicadas no âmbito educacional, tanto para estimar a habilidade dos estudantes, quanto para aprimorar as avaliações. Outras áreas também adotam a TRI: para estimação de grau de satisfação, avaliação de intenções comportamentais, avaliação da qualidade de vida, entre outros (Moreira Junior, 2010).

A escolha sobre qual modelo utilizar está diretamente ligada aos objetivos almejados. Para sistemas de recomendação de exercícios de programação entende-se que o mais apropriado é o modelo que apresenta o histórico das sub-habilidades dos usuários, pois de acordo com essas informações é possível lhe indicar problemas compatíveis, nem tão fáceis, nem tão difíceis, de forma a desafiá-los.

Como trabalho futuro pretende-se investigar os modelos M-ERS, Elo e TRIM com o objetivo de propor um modelo que agregue os pontos positivos de cada um. Acredita-se que um modelo híbrido, que contemple essas abordagens, poderá ser aplicado, de forma mais precisa, em sistemas de recomendação e na avaliação das habilidades e do processo de desenvolvimento de estudantes que utilizam plataformas *online* para resolver problemas de programação.

## Agradecimentos

Universidade Federal do Rio Grande (FURG) e Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSul). O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

## Referências

Andrade, D., & Justino, G. (2007). Software para avaliação de aprendizagem utilizando a teoria da resposta ao item. In *XIII Workshop sobre Informática na Escola.* Rio de Janeiro–RJ. doi: 10.5753/cbie.wie.2007 [GS Search]

Andrade, D., Tavares, H. R., & Cunha, V. R. (2000). *Teoria da resposta ao item: conceitos e aplicações*. São Paulo: ABE - Associação Brasileira de Estatística. [GS Search]

Araujo, E., Andrade, D., & Bortolotti, S. (2009). Teoria da resposta ao item. *Revista da Escola de Enfermagem da USP*, *43*, 1000-1008. doi: 10.1590/S0080-62342009000500003 [GS Search]

Baker, F. (2001). *The basics of item response theory* (2nd ed.). Washington: ERIC. [GS Search]

Bez, J. L., Tonin, N. A., & Rodegher, P. R. (2011). *URI Online Judge.* Retrieved from https://www.urionlinejudge.com.br/

Bez, J. L., Tonin, N. A., & Rodegheri, P. R. (2014). URI Online Judge Academic: A tool for algorithms and programming classes. In *9th international conference on computer science education* (p. 149-152). doi: 10.5753/wei.2015.10235 [GS Search]

Braga, B. (2015). *Teoria da resposta ao item: o uso do modelo de Samejima como proposta de correção para itens discursivos*. UnB, Brasília.

Chalmers, R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29. doi: 10.18637/jss.v048.i06 [GS Search]

Elo, A. (1978). *The rating of chessplayers, past and present*. London: Batsford. [GS Search]

França, L. (2020). *Competências e habilidades no ensino:o que são e como aplicá-las?* Retrieved from https://www.somospar.com.br/competencias-e-habilidades/

Moreira, G. L., Holanda, W., Coutinho, J. C., & Chagas, F. (2018). Desafios na aprendizagem de programação introdutória em cursos de TI da UFERSA, campus Pau dos Ferros: um estudo exploratório. *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA*(2). [GS Search]

Moreira Junior, F. (2010). Aplicações da teoria da resposta ao item (TRI) no Brasil. *Revista Brasileira de Biometria*, *28*(4), 137–70. [GS Search]

Nojosa, R. T. (2002). Teoria da Resposta ao Item (TRI): modelos multidimensionais. *Estudos em Avaliação Educacional*(25), 123–166. doi: 10.18222/eae02520022193 [GS Search]

Oliveira, L. (2017). *O desempenho em matemática do enem de 2012 em Luis Eduardo Magalhães (BA), na teoria de resposta ao item*. UFT, Arraias.

Park, J., Cornillie, F., Van der Maas, H., & Van Den, N. (2019). A multidimensional IRT approach

for dynamically monitoring ability growth in computerized practice environments. *Frontiers in psychology*, *10*. doi: 10.3389/fpsyg.2019.00620 [GS Search]

Pasquali, L. (2018). *Tri–teoria de resposta ao item: Teoria, procedimentos e aplicações*. Curitiba: Appris. [GS Search]

Pelánek, R. (2016). Applications of the elo rating system in adaptive educational systems. *Computers Education*, *98*, 169-179. doi: 10.1016/j.compedu.2016.03.017 [GS Search]

Perrenoud, P., & Magne, B. (1999). *Construir as competências desde a escola*. Porto Alegre: Artmed. [GS Search]

Prisco, A., Santos, R., Botelho, S., Tonin, N., & Bez, J. (2018). A multidimensional Elo model for matching learning objects. In *2018 IEEE Frontiers in Education Conference (FIE)*. [GS Search]

Reckase, M. D. (2006). Multidimensional item response theory. *Handbook of statistics*, *26*, 607–642. doi: 10.1016/S0169-7161(06)26018-8 [GS Search]

Robins, A. (2010). Learning edge momentum: A new account of outcomes in CS1. *Computer Science Education*, *20*(1), 37–71. doi: 10.1080/08993401003612167 [GS Search]

Soares, T. M., Souza, R. C., & Pereira, V. R. (2004). Métodos alternativos no critério Brasil para construção de indicadores sócio-econômicos: Teoria da resposta ao item. *XXXVI Simpósio Brasileiro de Pesquisa Operacional*, 35. [GS Search]

Tavares, C. (2014). *A teoria de resposta ao item na avaliação em larga escala: Um estudo sobre o exame nacional de acesso ao mestrado profissional em matemática em rede nacional-profmat*. IMPA, Rio de Janeiro.

Wasik, S., Antczak, M., Badura, J., Laskowski, A., & Sternal, T. (2018). A survey on online judge systems and their applications. *ACM Computing Surveys (CSUR)*, *51*(1), 1–34. doi: 10.1145/3143560 [GS Search]