Um estudo comparativo entre dois modelos que estimam a habilidade dos estudantes: ELO e Teoria de Resposta ao Item

Fabiana Zaffalon[1], André Prisco[1], Ricardo Souza[2], Rafael Penna[1], Jean Luca Bez[3], Neilor Tonin[3], Silvia Botelho[1]

[1]Programa de Pós-Graduação em Educação em Ciências
Universidade Federal do Rio Grande (FURG) – Rio Grande – RS – Brasil

[2]Programa de Pós-Graduação em Modelagem Computacional
Universidade Federal do Rio Grande (FURG) – Rio Grande – RS – Brasil

[3]Universidade Regional Integrada do Alto Uruguai e das Missões
Erechim – RS – Brasil

{fabinhazaffalon,prisco.c3,rapennas,silviacb.botelho}@gmail.com,

rcrdsou@hotmail.com, {bez, neilor}@urionlinejudge.com.br

Abstract. *Currently, there are many online platforms that offers programming exercise libraries where evaluation occurs automatically. The present work presents an analysis of two models that aims to estimate the students' ability: ELO and TRI Theory. ELO was developed to classify players through game history, and TRI estimates skill through a set of responses given to a set of items. For the application of the models we use a database made available by an Online Judge platform. The results show us differences between the models in relation to the estimated abilities, differences that we believe are related to the way in which each model estimates the parameters.*

Resumo. *Atualmente, há muitas plataformas online que disponibilizam bancos de exercícios de programação, onde a avaliação destes ocorre de forma automática. O presente trabalho apresenta uma análise de dois modelos que tem por objetivo estimar a habilidade dos estudantes: ELO e Teoria TRI. O ELO foi desenvolvido para classificar jogadores, através do histórico de jogo, e a TRI estima a habilidade através de um conjunto de respostas dadas a um conjunto de itens. Para aplicação dos modelos utilizamos uma base de dados disponibilizada por uma plataforma Online Judge. Os resultados obtidos nos apontam diferenças entre os modelos em relação as habilidades estimadas, diferenças que acreditamos estar relacionadas à forma com que cada modelo estima os parâmetros.*

1. Introdução

Educadores buscam o melhoramento dos métodos para avaliar a proficiência e habilidade dos estudantes. Nos últimos anos os métodos de avaliação educacional com maior destaque são aqueles que visam apresentar dados precisos sobre a construção das competências, sejam de estudantes presenciais ou à distância através do uso de plataformas online que utilizam avaliações automáticas. Assim, o propósito deste trabalho é analisar

DOI: 10.5753/cbie.sbie.2019.459

dados obtidos a partir de dois métodos que estimam a habilidade dos aprendizes: ELO e Teoria de Resposta ao Item (TRI).

O ELO [Elo 1978], originalmente criado para avaliar habilidades de jogadores de xadrez, objetiva classificar os jogadores através dos seus históricos de jogo. É um sistema de classificação estatística que pode calcular valores relativos ao nível de habilidade para competidores ou máquinas em jogos competitivos. Nesse trabalho utilizamos um ELO adaptado [A. Prisco and Bez 2017] e [A. Prisco and Bez 2018], enquanto o modelo original lida com a relação de um jogador para outro jogador, essa abordagem adapta a métrica para relacionar o aprendiz ao problema com o qual interage, ou seja, um aprendiz compete com um problema e ambos recebem valores de ELO e os mesmos são atualizados no final de cada ”partida”ou no final da resolução do problema.

A TRI, é um conjunto de modelos matemáticos, que busca representar a probabilidade de um indivíduo acertar a resposta referente a um item em função dos parâmetros do mesmo e da(s) habilidade(s) do indivíduo. Tal relação, entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes, se expressa de forma que quanto maior a habilidade do indivíduo, maior será a probabilidade de acerto no item [D.F. Andrade and Valle 2000].

Em nossos experimentos, utilizamos um banco de dados de uma plataforma $OnlineJudge$ do Brasil [A. Prisco and Bez 2017], que contém problemas de programação onde os usuários (aprendizes) da plataforma resolvem esses problemas e recebem um $feedback$ de acerto ou erro. Esse experimento, com ELO e TRI, tem por finalidade entender as relações entre os dois modelos.

Nas seções seguintes apresentamos o referencial teórico que embasou nossa pesquisa, a metodologia utilizada para a execução dos experimentos e por fim, os resultados obtidos.

2. Referencial Teórico

2.1. Teoria de Resposta ao Item (TRI)

Teoria de Resposta ao Item (TRI) é uma abordagem utilizada para avaliação de medidas padronizadas e vem gradativamente sendo inserida no espaço da educação por ser considerada um instrumento importante no processo quantitativo de avaliação educacional, pois a mesma permite a construção de escalas de habilidades calibradas [D.F. Andrade and Valle 2000] e [Baker 2001]. A TRI compreende modelos matemáticos para obter as estimativas das habilidades, também chamada de $Theta(\theta)$, através das respostas dadas ao conjunto de itens em determinada área de conhecimento a ser avaliada. A estimativa da habilidade tem relação com a probabilidade de o indivíduo acertar o item, considerando um ou mais parâmetros. Dessa forma, quanto maior a habilidade do indivíduo, maior será a probabilidade de acerto ao item.

Existem vários modelos propostos da TRI, que 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) [D.F. Andrade and Valle 2000] e [Baker 2001]. Para itens dicotômicos, há 3 modelos que diferem entre si pela quantidade de parâmetros utilizados para descrever o item. São eles: Modelo de Rasch (considera somente a dificuldade do item), Modelo de 2

parâmetros (considera a dificuldade do item e a discriminação do mesmo) e Modelo de 3 parâmetros (considera a dificuldade, a discriminação e a probabilidade de acerto ao acaso) [D.F. Andrade and Valle 2000] e [Baker 2001].

Segundo [Baker 2001] o mais utilizado é o Modelo Logístico de 3 parâmetros, cuja equação é apresentada na Equação 1. Calcula a probabilidade do indivíduo com habilidade (θ) 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].

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

O parâmetro de discriminação do item (a) indica o quanto um item distingue os indivíduos com diferentes níveis de habilidade, ao imaginarmos 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 alunos com habilidades distintas responderem corretamente o item. Itens com maiores valores de a_j fornecem melhores discriminações[Tavares 2014].

A dificuldade do item (b) se dá na mesma escala da habilidade, ou seja, a habilidade que o indivíduo tem que possuir para uma dada probabilidade de responder corretamente item [Oliveira 2017] e [E.A.C. Araujo and Bortolotti 2009].

O parâmetro de acerto ao acaso (c) 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 assume valor 0 (zero)[D.F. Andrade and Valle 2000] e [Oliveira 2017].

Na TRI, a habilidade tem relação com a probabilidade do indivíduo responder corretamente um item através de funções matemáticas. Tais funções precisam ser estimadas e são chamadas de ”Curva Característica do Item”(CCI). A CCI representa a relação existente entre a habilidade e o desempenho nos itens. Mais informações sobre a CCI, podem ser encontrada em [D.F. Andrade and Valle 2000], [Andrade and Justino 2007] e [Baker 2001].

Os modelos para itens não dicotômicos são modelos que analisam itens abertos (resposta livre) e itens de múltipla escolha que são avaliados de forma graduada. Um dos modelos para esse tipo de item é o Modelo de Resposta Gradual, de Samejima, que é uma generalização do modelo logístico de 2 parâmetros e assume que as categorias de respostas podem ser ordenadas entre si, como uma escala de Likert, que consiste em uma série de 5 proposições [Braga 2015]. Demais modelos são encontrados em [D.F. Andrade and Valle 2000] e [Braga 2015].

Um dos principais passos da TRI é a estimação dos parâmetros dos itens (calibração) e das habilidades (traços latentes) e essa etapa necessita de recursos computacionais. Os modelos mais utilizados são o Método Bayesiano e o de Máxima Verossimilhança Marginal [D.F. Andrade and Valle 2000], [E.A.C. Araujo and Bortolotti 2009] e [Chalmers 2012].

2.2. Sistema de Classificação ELO

O sistema de classificação ELO foi proposto para avaliar jogadores de xadrez. Cada jogador recebe uma classificação inicial (um ELO) e, à medida que vai participando dos jogos, essa classificação vai sendo atualizada de acordo com os resultados. O modelo trabalha em função da expectativa e do resultado. Se o resultado atende à expectativa (por exemplo, um jogador com maior habilidade vence um jogador com menor habilidade) os valores de ELO recebem pequenas atualizações. Já em caso contrário, a atualização é maior [Elo 1978], [Pelánek 2016] e [A. Prisco and Bez 2017]. Em [A. Prisco and Bez 2018] é apresentada uma adaptação da técnica ELO semelhante à usada na escolha de oponentes em torneios de xadrez ou partidas $online$. A expectativa de que i ganhe de j $(R_{ij} = 1)$ é dado pela Equação 2 [A. Prisco and Bez 2017]:

$$P(R_{ij} = 1) = \frac{1}{1 + 10^{\frac{\theta_j - \theta_i}{400}}} \tag{2}$$

Onde $R = \{0, 1\}$ é o conjunto de resultados de um jogo: 1 (ganhar) e 0 (perder). Dado um jogo entre o jogador i e o jogador j, com um ELO θ_i e θ_j, respectivamente. Ao final da partida novos ELOs são calculados de acordo com as expectativas dos resultados, os ELOs anteriores e uma constante k. Quanto maior o k, maior é a mudança do ELO, conforme Equação 3 [A. Prisco and Bez 2017] e [A. Prisco and Bez 2018]:

$$\theta_i = \theta_i + k(R_{ij} - P(R_{ij} = 1)) \tag{3}$$

O ELO lida com um jogador relacionado com outro jogador, porém nossa abordagem relaciona um usuário a um problema. Assim, para cada submissão, consideramos um jogo, ou seja, um ”duelo”entre o jogador e o problema. Em cada início de submissão, o algoritmo armazena o ELO do problema e do aluno. Após a avaliação do “duelo”, os ELOs, tanto do usuário quanto do problema, são atualizados.

O ELO pode ser utilizado na educação, se julgarmos que uma tentativa de solução tem uma relação entre item e aluno, estimando a habilidade do aluno e a dificuldade do item [Pelánek 2016]. A estimação acontece de maneira contínua, pois a atualização da classificação acontece ao término de cada evento.

Segundo [Pelánek 2016], existe uma relação entre o sistema de classificação ELO e a TRI de um parâmetro. 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 que o ELO foi implementado para rastrear as mudanças nos níveis de habilidades.

2.3. $OnlineJudges$

$OnlineJudges$ são ambientes virtuais de aprendizagem que possuem um repositório de problemas de programação onde os aprendizes (estudantes) ou usuários podem competir resolvendo tais problemas [A. Prisco and Bez 2017].

Professores e estudantes incorporaram o uso dessa plataforma como uma ferramenta didática. Para resolver o problema, os estudantes devem elaborar um algoritmo (programa) em uma das linguagens de programação aceitas pela plataforma. A plataforma tem um sistema automático de avaliação que faz a análise do programa submetido

e dá um $feedback$ ao estudante. Os tipos de $feedback$ são: aceito (problema aceito sem erro) ou de erro (compilação, execução, apresentação, tempo de execução, tempo excedido, falha de comunicação com o servidor ou resposta errada). A ferramenta permite que o usuário resolva os problemas aleatoriamente e submeta uma quantidade indeterminada de resolução para o mesmo problema, mesmo que a resposta esteja correta.

3. Metodologia

Utilizamos a base de dados disponibilizada pelo $OnlineJudge$ no Brasil, com identificadores fictícios dos usuários. Foram realizados 3 experimentos: (1) ELO e TRI modelo 2PL, considerando todas as submissões; (2) ELO e TRI 2PL considerando apenas uma submissão por problema e (3) ELO e TRI Modelo de Resposta Gradual. Nos 3 experimentos foram eliminados os dados replicados, os itens/problemas em que todos usuários acertaram ou erraram, e os usuários que submeteram uma grande quantidade de soluções em um espaço de tempo muito pequeno. Para facilitar o processamento, transformamos os dados textuais que recebemos do $OnlineJudge$, em dados numéricos: os problemas com soluções corretas, que foram aceitos, receberam o valor 1 (um) e os que não foram aceitos, receberam o valor 0 (zero).

O primeiro passo foi atribuir um valor de ELO inicial para cada usuário e para cada problema. Nesta fase inicial, todos os usuários recebem um ELO (1100), conforme [A. Prisco and Bez 2017]. Para a atribuição de ELO dos problemas, usamos como heurística o nível de dificuldade indicado pelo autor do problema da plataforma $OlineJudge$.

Com os ELOs iniciais atribuídos, o algoritmo simula cada submissão realizada em ordem cronológica. Através do $feedback$ de cada envio, os ELOs são atualizados. No final do processo, temos os ELOs finais de cada usuário e de cada problema. Os valores no final do processo são comparados com os valores pré-processados. Se a diferença entre os valores é substancial isso indica que os ELOs ainda não convergiram e o processo é repetido. Enquanto o processo converge, nós armazenamos os ELOs finais de cada problema. Quando nós conhecemos o ELO de cada problema, consideramos que este seja constante para o próximo passo. Uma vez que a etapa de convergência tenha sido concluída, os usuários novamente recebem ELOs médios e, para cada nova submissão, seus ELOs são atualizados e armazenados no banco de dados. Dessa forma, temos o histórico do ELO de cada usuário para cada submissão. Mais informações podem ser obtidas em [A. Prisco and Bez 2017] e [A. Prisco and Bez 2018].

Para a TRI analisamos o conjunto de respostas de cada problema, por usuário. Utilizamos o software RStudio e o pacote mirt que foi criado para estimar os parâmetros dos itens e a habilidade (traço latente), utilizando o método Máxima Verossimilhança Marginal [Chalmers 2012]. No experimento 1, consideramos somente as questões submetidas por todos os usuários e usuários que submeteram todos os itens, assim a base de dados foi composta por 1.006 usuários e 80 problemas. Mesmo que o usuário tenha submetido várias vezes a solução do mesmo problema, se em alguma dessas submissões a solução estivesse correta, consideramos como questão correta. Tabulamos as informações de forma a relacionar esses dados com as respostas, atribuindo 1 (um) para problema aceito e 0 (zero) para não aceito.

O modelo da TRI aplicado nesse experimento foi o de 2 parâmetros (2PL), modelo

que não estima o acerto ao acaso. O primeiro passo foi fazer a calibração e parametrização dos itens para obter o nível de dificuldade e a discriminação dos mesmos. Com os resultados dos parâmetros dos itens foi feita uma análise dos valores considerados críticos, o que proporcionou a validação dos itens considerados satisfatórios, segundo as características da TRI, bem como a exclusão ou reavaliação dos itens considerados fora destas condições [E.A.C. Araujo and Bortolotti 2009]. Todos os itens analisados foram satisfatórios. Após a estimativa dos parâmetros, o próximo passo foi obter o traço latente ou a habilidade dos usuários.

No experimento 2, foram considerados somente os problemas com apenas uma submissão pelos usuários, independente da resposta, totalizando 3.357 usuários e 100 itens. Novamente foram aplicados os modelos ELO e TRI 2PL, conforme apresentado acima.

Já no experimento 3, a base de dados foi composta por 4.626 usuários e 100 itens. Aplicamos os mesmos critérios de exclusão apresentado anteriormente e consideramos todos os problemas, mesmo que algum usuário não tenha submetido solução. O modelo ELO foi aplicado nessa base, seguindo a metodologia apresentada acima. O modelo da TRI aplicada foi o da Resposta Gradual, para esse modelo criamos escalas de submissões para cada usuário baseado na escala de Likert.

Foram identificadas até 167 submissões do mesmo usuário para determinado problema. Assim, criamos as seguintes escalas de acordo com o número de submissões: Escala 1 (uma submissão) com 326.687 ocorrências; escala 2 (de duas a cinco submissões) com 99.603 ocorrências; escala 3 (de seis a dez submissões) com 11.110 ocorrências; escala 4 (mais de dez submissões) com 3.060 ocorrências; e escala 5 (submeteu, uma ou mais vezes, mas não obteve acerto) com 12.089 ocorrências. Entre as escalas um e quatro, em alguma submissão a solução estava correta.

Com as escalas definidas, aplicamos a TRI Modelo de Resposta Gradual, utilizando o pacote mirt da linguagem R, para estimar os parâmetros dos itens e as habilidades dos usuários. De posse das habilidades encontradas nos três experimentos, realizamos a análise dos dados, discutidos na próxima seção.

4. Resultados e Discussões

No experimento 1, analisamos o histórico ELO de cada um dos usuários. O que percebemos foi a variação no valor ELO, tanto do usuário quanto do problema, que ocorre de acordo com as submissões, pois o ELO considera todas as submissões. Por exemplo, os problemas corretos na primeira submissão, tem seu ELO atualizado numa determinada escala, simultaneamente o ELO do usuário também é atualizado, diminuindo seu valor ELO ou a escala de dificuldade do problema. O contrário também acontece, usuários que submetem várias vezes, até acertar a solução do mesmo problema, tem o seu valor de ELO atualizado, porém numa escala menor, consequentemente o ELO daquele problema aumenta.

A Figura 1 ilustra o histórico de submissões de 4 usuários, escolhidos aleatoriamente, com diferentes perfis de variação na habilidade ELO. O eixo horizontal apresenta a quantidade de submissões realizadas pelos usuários. O eixo vertical apresenta o valor de ELO a cada submissão.

Tomaremos como exemplo um usuário (user 1) que submeteu 144 soluções, das quais 90 estavam corretas e 54 incorretas . Dos problemas que foram submetidos e estavam corretos, 73 foram submetidos uma vez, 2 problemas foram submetidos 5 vezes e demais problemas foram submetidos entre 2 e 4 vezes, obtendo o acerto somente na última submissão. Um desses problemas com 5 submissões até o acerto, foi submetido entre a quadragésima primeira e a quadragésima quinta submissão, o que justifica a queda do valor ELO no gráfico da Figura 1 entre as escalas 41 e 45. O mesmo evento ocorre com os demais usuários e intervalos onde o gráfico ilustra o valor de ELO diminuindo e aumentando.

Figura 1. Histórico ELO de submissões dos usuários

A maioria dos usuários ficou com ELO por volta de 1200, ou seja, a maioria teve aumento de ELO e os valores foram distintos, não houve ocorrência de mais de um usuário com o mesmo valor ELO. Não há uma dispersão muito grande entre os usuários devido ao nível de dificuldade dos problemas.

Na mesma base utilizada pelo ELO, foram estimados os parâmetros dos problemas e as habilidades dos usuários pelo modelo TRI. Porém, para que pudéssemos comparar o ELO com as habilidades da TRI, selecionamos o ELO final de cada usuário, ou seja, o ELO da última submissão de cada problema.

Na TRI, muitos usuários receberam o mesmo valor de habilidade, isso se deu porque os usuários submeteram várias soluções até obter o acerto da questão. Como neste neste caso de análise o foco era identificar se acertou ou errou o item, a quantidade de submissões não foi considerada, assim, a maioria dos usuários acabou acertando muitos itens. Para melhor analisar os resultados, foi feito um ranking de classificação, do 1º ao 385º. A Tabela 1 apresenta alguns dos rankings.

Tabela 1. Ranking TRI

Ranking	Quantidade de Usuários	Percentual de Usuários
1º	510	50,69%
2º	55	5,4671%
3º	1	0,0994%
21º	10	0,9949 %

Dos 1.006 usuários, 50,69%, que corresponde a 510, obtiveram o maior valor de TRI (0,727555877), ou seja, 510 ocuparam o primeiro lugar. Isso aconteceu porque

todos eles acertaram todas as questões. O segundo lugar, está ocupado por 5,46%, que corresponde a 55 usuários, com TRI igual a 0,58533256. Os mesmos erraram a mesma questão, obtendo acerto nas demais. Apenas um usuário (0,099%) ficou em terceiro lugar no ranking, com o mesmo número de acertos dos usuários do ranking 2, porém com TRI no valor de 0,5494. Os usuários do ranking 2 erraram uma questão cujo parâmetro de discriminação é menor que o parâmetro de discriminação do problema que o usuário que está em terceiro lugar errou, por isso se deu a diferença no valor de TRI. O mesmo ocorreu com os demais rankings e usuários, em que a dificuldade e as demais características das questões contribuem com o resultado. Dessa forma a habilidade é estimada de acordo com o parâmetro de discriminação dos itens que os usuários acertam ou erram, isto é, acertar a mesma quantidade de questões que outros usuários não significa que terão a mesma habilidade.

Na Tabela 2 relacionamos as habilidades dos dois modelos, ELO e TRI, dos mesmos usuários analisados no ELO do experimento 1. Tendo em vista que os valores dos ELOs variaram entre 555 e 1.475 e a TRI entre -3 e 0,50, a Tabela 2 mostra que nem sempre a relação dos valores das habilidades dos dois modelos são equivalentes. Ou seja, o usuário com maior ou menor ELO pode não ser o mesmo com maior ou menor habilidade TRI. Como exemplo, podemos observar os usuários 1 e 4, que apresentam ELOs altos e TRI baixos. Porém, se observarmos o usuário 2 percebemos habilidades altas nos dois modelos. Esse usuário é um dos que acertou todos os problemas, por isso recebeu o maior valor de TRI, mas não o maior valor de ELO. Essa diferença se justifica pela quantidade de vezes que o mesmo submeteu algumas soluções até acertá-las.

Tabela 2. Habilidades estimadas pelo ELO e TRI

idlUser	ELO	TRI
1	1272,18	-1,663830070
2	1461,77	0,49049927
3	740,088	-1,818720554
4	1424,5	-1,64809031

Normalizamos as habilidades dos dois modelos e calculamos os percentis da diferença destas habilidades, com o objetivo de dividir a amostra para melhor analisar a distribuição dos usuários, ilustrada na Figura 2.

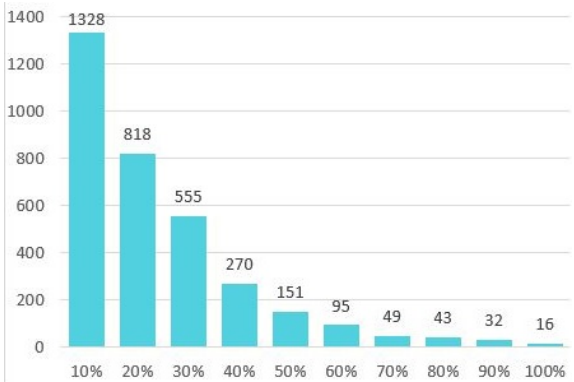

Figura 2. Percentis da diferença entre as habilidades (experimento 2)

De acordo com o gráfico da Figura 2, dos 3.357 usuários, 1.328 estão na faixa de 10%. Isso significa que esses usuários têm 10% de diferença nas habilidades estimadas pelo ELO e TRI. 818 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 grande diferença na estimativa das habilidades entre os modelos.

No experimento 3, as habilidades ELO variaram entre 555 e 1.474 e a na TRI Gradual as habilidades variaram entre -4 e +4. Da mesma forma que no experimento 2, cada usuário ficou com um valor único de TRI. Para melhor comparar as habilidades TRI e ELO, também normalizamos as habilidades dos dois modelos e calculamos os percentis das diferenças dos mesmos, conforme mostra a Figura 3.

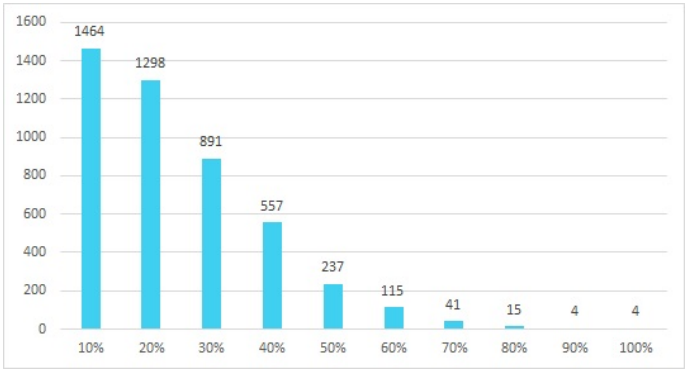

Figura 3. Percentis da diferença entre as habilidades (experimento 3)

De acordo com o gráfico da Figura 3, 1.464 usuários estão na faixa dos 10%, o que significa que esses usuários têm 10% de diferença nas habilidades dos modelos ELO e TRI e, assim, sucessivamente.

Podemos observar que mesmo alterando o modelo da TRI, a maioria dos usuários ficou com uma diferença de 10% nas habilidades, o mesmo ocorreu no experimento 2. Comparando os dois modelos da TRI, verificamos que os gráficos dos dois modelos, (Figura 2 e Figura 3) estão semelhantes, porém no modelo gradual (Figura 3) há um número menor de usuários que estão nos maiores percentis, isto é, há menos usuários com maiores diferenças entre as habilidades estimadas.

Na análise feita no experimento 1, nossa hipótese é que essa diferença possa estar relacionada à forma com que cada modelo estima as habilidades. Os modelos da TRI baseiam-se em suposição de uma habilidade constante ou fixa, não leva em consideração o desempenho dos usuários ao longo do tempo, ou seja, foi descartado o número de vezes que o usuário submeteu as soluções até acertar o item ou desistir de acertar, enquanto o modelo de classificação ELO acompanha a habilidade de mudança do usuário e adapta o grau de dificuldade dos problemas e habilidade dos mesmos, considerando todas interações entre o usuário e os problemas.

No segundo experimento, houve uma diferença bem menor entre os valores das habilidades entre os dois modelos. Nossa hipótese é que essa diferença diminuiu porque no ELO não foram consideradas as interações dos usuários com o problema, uma vez que cada problema teve apenas uma submissão.

Já o terceiro experimento também apresentou um resultado diferente de TRI para cada usuário e, ao comparar com o ELO, os valores das habilidades se aproximaram mais do que os da TRI de 2 parâmetros. Acreditamos que essa diferença tenha ocorrido pelo fato de termos criado escalas referentes ao número de submissões, isto é, o número de submissões não foi descartado, conforme ocorreu no experimeto 2. Esta foi uma tentativa de analisar se, de alguma forma, este modelo da TRI mensura a evolução do usuário, considerando a quantidade de submissões até o acerto, como ocorre com o ELO.

Os resultados obtidos mostram que há diferença entre as habilidades estimadas nos dois modelos. Ainda não conseguimos encontrar um modelo TRI que seja similar ao ELO quando aplicados na avaliação de resolução de algorítmos. Como próximos passos, planejamos pesquisar os modelos de TRI temporal para estudar uma forma de estimar as habilidades, ao longo do tempo, dos estudantes que utilizam plataformas $online$ de programação, unificando os dois modelos. Acreditamos que um modelo híbrido, que contemple ambas abordagens poderá ser aplicado na avaliação da habilidade e do processo de desenvolvimento de estudantes neste tipo de plataforma.

Referências

A. Prisco, R. Santos, S. B. N. T. and Bez, J. (2017). Using information technology for personalizing the computer science teaching. In *2017 IEEE Frontiers in Education Conference (FIE)*.

A. Prisco, R. Santos, S. B. N. T. and Bez, J. (2018). A multidimensional elo model for matching learning objects. In *2018 IEEE Frontiers in Education Conference (FIE)*.

Andrade, D. and Justino, G. (2007). Software para avaliação de aprendizagem utilizando a teoria da resposta ao item. In *Anais do XXVII Congresso da SBC)*.

Baker, F. (2001). *The basics of Item Response Theory*. ERIC.

Braga, B. (2015). Teoria da resposta ao item: o uso do modelo de samejima como proposta de correção para itens discursivos. Master's thesis, UnB.

Chalmers, R. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48:1–29.

D.F. Andrade, H. T. and Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. In *Simpósio Nacional de Probabilidade e Estatística (SINAPE)*.

E.A.C. Araujo, D. and Bortolotti, S. (2009). Teoria de resposta ao item. *Revista da Escola de Enfermagem da USP*, 43:1000–1008.

Elo, A. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Oliveira, L. (2017). O desempenho em matemática do enem de 2012 em luis eduardo magalhães (ba), na teoria de resposta ao item. Master's thesis, UFT.

Pelánek, R. (2016). Applications of the elo rating system in adaptive educational systems. *Computers Education*, 98:169–179.

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. Master's thesis, IMPA.