Explorando Teoria de Resposta ao Item na Avaliação de Pensamento Computacional: um Estudo em Questões da Competição Bebras

Ana Liz Souto O. Araujo[1,2], Wilkerson L. Andrade[1], Dalton D. S. Guerrero[1], Monilly Ramos A. Melo[3], Isabelle Maria Lima de Souza [1]

[1]Laboratório de Práticas de Software (SPLab)
Universidade Federal de Campina Grande (UFCG)

[2]Departamento de Ciências Exatas (DCX)
Universidade Federal da Paraíba (UFPB)

[3]Laboratório de Neuropsicologia e Inovação Tecnológica
Universidade Federal de Campina Grande (UFCG)

analiz@dcx.ufpb.br, {wilkerson,dalton}@computacao.ufcg.edu.br

monilly.ramos@gmail.com, isabellemaria@copin.ufcg.edu.br

Abstract. *Evaluating Computational Thinking (CT) is a research gap, mainly due to the complexity of assessing cognitive abilities. Although the Bebras Challenge is an initiative to stimulate PC, we still do not know its reliability to measure CT. In this exploratory study, we investigated whether the tasks produced for the Bebras Challenge can be used to assess CT using the Item Response Theory. The results showed that the tasks have not only moderate and high discrimination but also easy and medium difficulty level. However, the analyzed tasks do not present adequate level of reliability to be used as a measuring instrument.*

Resumo. *Avaliar Pensamento Computacional (PC) é uma questão de pesquisa em aberto, principalmente pela complexidade de se avaliar habilidades cognitivas. Embora a competição Bebras seja uma iniciativa para estimular PC, ainda não sabemos sua adequação para mensurar PC. Neste estudo exploratório, investigamos se as questões produzidas para a competição Bebras podem ser usadas para avaliação de PC usando como método a Teoria de Resposta ao Item para estimar os parâmetros dos itens. Os resultados apontaram que as questões analisadas apresentam discriminação moderada e alta, e dificuldade fácil e moderada. Entretanto, o conjunto de questões analisado ainda não apresenta nível de confiabilidade adequado para ser usado como instrumento avaliativo.*

1. Introdução

Nos últimos anos, diversos estudos se propõem a estimular o Pensamento Computacional (PC), todavia a forma de avaliação ainda é um tópico que requer mais investigação [Avila et al. 2017, Araujo et al. 2016]. A dificuldade em avaliar PC recai na própria complexidade que se constitui avaliar construtos, como por exemplo, a inteligência. Construtos são habilidades cognitivas que não podem ser mensurados di-

DOI: 10.5753/cbie.sbie.2018.665

retamente [Hutz et al. 2015]. Além disso, existe um obstáculo a mais para se propor soluções para avaliar PC: a falta de um consenso sobre a definição operacional [Román-González et al. 2016].

As propostas de avaliação concentram-se em medir PC durante experiências de ensino de algoritmos e programação [Avila et al. 2017]. Nesses casos, os ambientes de programação são usados tanto no contexto de ensino como no contexto avaliativo, bem como ferramentas auxiliares que utilizam os artefatos produzidos pelos alunos. Dr. Scratch [1] é uma das ferramentas de análise empregadas quando se utiliza a linguagem de blocos Scratch. Por outro lado, outras propostas de avaliação de artefatos de código têm sido propostas, como o uso de mapas auto-organizáveis a partir de técnicas de aprendizado de máquina [Barcelos et al. 2017]. As limitações dessas abordagens recaem na necessidade de ensinar algoritmo e programação para então poder mensurar PC.

Jogos e testes são abordagens que podem ser projetados de forma que não exijam conhecimento em programação para avaliar PC e que se baseiem em escore como medida de avaliação. Raabe *et al* apresentaram um jogo *puzzle* que utiliza o desempenho na resolução das atividades de cada fase para apresentar um escore de PC por jogador [Raabe et al. 2017]. Por outro lado, o *Computational Thinking Test* é um teste de PC para adolescentes espanhóis de 12 a 14 anos [Román-González et al. 2016]. O teste, embora não exija conhecimentos em programação, foca em conceitos computacionais relacionados à programação, como *loops*, estruturas de repetição e decisão por meio de linguagem de blocos nas questões.

Bebras [2] é uma iniciativa internacional para disseminar PC que, embora não tenha o objetivo testar conhecimentos, é possível que possa ser usado também para avaliar PC futuramente [Dagiene and Stupuriene 2016, Araujo et al. 2017]. A competição Bebras acontece anualmente por meio de uma prova contendo 15 a 18 questões que não necessitam de conhecimento prévio em programação para serem respondidas. A pontuação da competição é feita de acordo com a dificuldade das questões, classificadas pelos organizadores em três níveis: fáceis, médias e difíceis. Dessa forma, a pontuação final dos competidores é calculada baseada no escore de acertos e erros das questões em cada nível de dificuldade.

Estudos anteriores apontaram que o método de pontuação baseado em escore empregado em competições Bebras não se mostrou adequado como forma de avaliação de PC, mas que as questões elaboradas podem ser promissoras [Araujo et al. 2017]. Neste estudo, continuamos a investigar se as questões produzidas para a competição Bebras podem ser usadas para avaliação de PC explorando outro recurso estatístico de análise de itens. Mais especificamente, o objetivo deste estudo é explorar parâmetros psicométricos estimados pela Teoria de Resposta ao Item (TRI) para avaliar as questões elaboradas pelo Bebras. Dentre os parâmetros estimados, comparamos a classificação de dificuldade das questões fornecida pelos organizadores do Bebras com o resultado apontado pela TRI. Para atingir esses objetivos, coletamos dados dentro do contexto do primeiro semestre de cursos de Computação. Aplicamos uma prova Bebras para uma amostra de 126 alunos ingressantes em cursos de Computação de duas universidades. Os parâmetros da TRI para os itens foram estimados a partir do banco de respostas obtido e os resultados são

[1] http://www.drscratch.org/
[2] www.bebras.org

apresentados e discutidos.

O artigo segue organizado da seguinte forma: a Seção 2 apresenta conceitos sobre Pensamento Computacional, detalha a iniciativa Bebras e discute os trabalhos relacionados. A Seção 3 apresenta brevemente os conceitos elementares da TRI. A Seção 4 descreve a metodologia adotada. Os resultados e as discussões são apresentadas nas Seções 5 e 6, respectivamente. Por último, as considerações finais estão descritas na Seção 7.

2. Pensamento Computacional

Em uma definição mais ampla, PC pode ser compreendido como um processo de reconhecer aspectos da Computação em atividades do cotidiano, aplicar técnicas e ferramentas da computação para entender e raciocinar sobre aspectos naturais e sociais de sistemas e processos computacionais [Csizmadia et al. 2015]. Assim, PC envolve resolver problemas, projetar sistemas e entender o comportamento humano por meio dos conceitos fundamentais da Computação [Wing 2006]. Em uma definição mais operacional, PC permite que as pessoas lidem com um problema identificando elementos essenciais, quebrando-o em partes menores, mais fáceis de manejar e planejem uma sequência de passos para solucioná-lo [Csizmadia et al. 2015]. Assim, PC é processo cognitivo que envolve raciocínio usando habilidades computacionais pelo qual os problemas são resolvidos.

2.1. Bebras

Bebras é uma comunidade internacional de Educação em Computação originada na Lituânia que tem o objetivo de promover PC. A comunidade se reúne uma vez ao ano para elaborar questões que são usadas posteriormente na competição anual que ocorre em escolas de 40 países [Dagiene and Stupuriene 2016]. O nome Bebras foi escolhido para competição porque o castor (tradução de *bebras* do lituano) é um tipo de roedor considerado trabalhador, inteligente e que busca objetivos. Assim, as questões elaboradas envolvem castores enfrentando um problema que precisa ser solucionado explorando competências de PC.

As questões são elaboradas de tal forma que incluem um conceito de computação, mas que não exigem conhecimento prévio em Computação ou programação para resolvê-la [Dagiene and Stupuriene 2016]. As questões são produzidas em inglês e ficam armazenadas em um banco de questões. Os organizadores de cada país participante podem escolher questões desse banco, traduzir para sua língua nativa e classificar as questões quanto ao nível dificuldade: fácil, médio e difícil para cada ano escolar participante da competição.

Dependendo no nível de dificuldade da questão, o aluno recebe uma pontuação diferente quando acerta ou quando erra a questão. Quando o aluno acertar uma questão fácil, ele ganha 6 pontos e se errar, não perde pontos. Já se o aluno acertar uma questão de dificuldade média, ele ganha 9 pontos e se errar, perde 2 pontos. Por último, se o aluno acertar uma questão difícil, ele ganha 12 pontos, e se errar, perde 4 pontos. As questões deixadas sem resposta não são contabilizadas, ou seja, o aluno nem perde nem ganha pontos.

2.2. Trabalhos Relacionados

Na literatura existem trabalhos que abordam diferentes métodos para avaliar PC. A avaliação por meio de artefatos de código é uma estratégia útil para ser usada junto com

oficinas e cursos de programação. Barcelos *et al.* (2017) propuseram o emprego de técnicas de aprendizado de máquina em código Scratch para avaliar o desenvolvimento de PC. Eles conseguiram relacionar as rubricas de habilidades de PC no código aos agrupamentos dos diferentes tipos jogos produzidos. Assim, eles viabilizam uma forma automatizada de análise de código para identificar jogos que exploram habilidades de PC semelhantes. Apesar das vantagens de usar uma forma automatizada de avaliação, são necessárias outras maneiras de avaliar PC sem vincular obrigatoriamente ao ensino de programação.

O uso de jogos projetados para avaliar PC é uma estratégia promissora quando se utiliza técnicas válidas para nortear tanto o desenvolvimento como o método de avaliação. Raabe *et al.* (2017) propuseram um jogo do tipo *puzzle* baseado em atividades de testes de QI e Questões do Programa de Enriquecimento Instrumental de Feuerstein para avaliar PC. A avaliação é realizada por meio de escores calculados em cada fase durante o jogo, como quantidade de interação, de dicas consumidas e de tempo transcorrido. Portanto, esse jogo pode fornecer uma forma de mensurar PC sem a necessidade de ensino de programação, mas se encontra em fase de validação.

Testes que permitem a avaliação de PC sem requerer conhecimento prévio em programação são úteis pela possibilidade de serem aplicados em larga escala. Dentre os poucos testes validados nessa vertente, destacamos o *Computational Thinking Test* (CTt) que avalia PC em alunos espanhóis de 12 a 14 anos [Román-González et al. 2016]. O CTt possui 28 itens objetivos e envolve conceitos computacionais, como *loops*, estruturas de repetição e decisão por meio de linguagem de blocos. Portanto, o CTt pode ser considerado um teste que avalia PC por meio de código em linguagem de blocos. Além disso, o teste fornece uma avaliação sumativa, baseado em escore.

3. Teoria de Resposta ao Item

Nesta seção apresentamos conceitos elementares da TRI para itens dicotômicos, ou seja, itens que podem ser corrigidos como certo ou errado. Maior aprofundamento no tema pode ser encontrado em [Hutz et al. 2015, Baker 2001].

TRI pode ser compreendido como um conjunto de modelos matemáticos elaborados para estimar o nível de habilidade, chamada de Theta (Θ), de um sujeito baseado em itens de um instrumento [Baker 2001]. As principais vantagens da TRI frente as outras teorias de mensuração consistem em avaliar cada item do instrumento separadamente, fornecendo parâmetros individuais, e não utilizar o escore como medida de avaliação, e sim utilizar a dificuldade dos itens e outros parâmetros, como discriminação, para estimar a habilidade do sujeito [Hutz et al. 2015].

De acordo com a TRI, é possível examinar até três parâmetros em um instrumento projetado para avaliar um construto [Baker 2001]. O parâmetro de discriminação descreve o quão bem o item é capaz de diferenciar sujeitos com níveis de habilidades (Θ) próximas. Portanto, um item com alta discriminação consegue detectar pequenas variações no nível da habilidade (Θ) medida. Já o parâmetro de dificuldade estipula a probabilidade de 50% do sujeito em acertar o item baseado no seu nível de aptidão e posiciona o item no intervalo da escala de habilidade (Θ). Por exemplo, sujeitos com baixa habilidade têm maior probabilidade de acertar um item fácil, enquanto que sujeitos com alta habilidade têm maior probabilidade de acertar um item difícil. Por último, o parâmetro de acerto

ao acaso estipula a probabilidade de um sujeito acerta o item ao acaso (por chute), sem possuir habilidade (Θ) necessária para acertá-lo.

O modelo logístico de três parâmetros (3PL) é usado para analisar a discriminação, dificuldade e probabilidade de acerto ao acaso dos itens de um instrumento quando os dados se ajustam a esse modelo. A equação fundamental do modelo 3PL é a Equação 1, que descreve a probabilidade de um sujeito qualquer com habilidade Θ responder corretamente a um item j, baseado no índice de discriminação a, de dificuldade b e de acerto ao acaso c do item j [Baker 2001]. O intervalo teórico da habilidade (Θ) é infinita, mas na prática, analisa-se o intervalo de -3 a 3 e se representa no gráfico o intervalo de -4 a 4.

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

A Curva Característica do Item (CCI) permite uma visualização gráfica dos parâmetros de um item. A CCI é plotada entre o eixo x representando a habilidade (Θ), geralmente variando de -4 a 4, e o eixo Y representando a probabilidade de acerto, variando de 0 a 1. Quanto mais próximo do formato de "S" a CCI de um item for, maior o nível de discriminação. Os itens mais fáceis são posicionados mais à esquerda no eixo x, enquanto que os itens mais difíceis são posicionados mais à direita [Hutz et al. 2015].

4. Metodologia

Para atingir os objetivos deste estudo, apresentamos os participantes, o instrumento e os procedimentos realizados na análise dos dados coletados.

4.1. Participantes

Os participantes foram 126 alunos ingressantes em cursos superiores de Computação de duas universidades públicas em 2017. Essa amostra foi selecionada por conveniência, pois todos os alunos regularmente matriculados no semestre letivo foram convidados a participar, e apenas os que aceitaram responderam o instrumento.

4.2. Instrumento

Utilizamos como instrumento uma prova Bebras aplicada no Reino Unido e disponibilizado pela organização [Blokhuis et al. 2014]. Como ainda não existem competições Bebras no Brasil, não há provas em português, e por isso, traduzimos as questões do inglês. Esse processo de tradução contou com a ajuda de colaboradores externos que revisaram as questões traduzidas.

O instrumento aplicado contém 15 questões. Selecionamos essas questões tomando como base aquelas direcionadas ao público mais velho (acima de 16 anos) da competição. Os organizadores Bebras do Reino Unido classificaram as questões de acordo com o nível de dificuldade. Assim, a prova que utilizamos contém cinco questões fáceis, cinco médias e cinco difíceis segundo essa classificação [Blokhuis et al. 2014]. Na Tabela 1 apresentamos as questões organizadas com o identificador do item, seguido pelo nome da questões e a classificação de dificuldade fornecida pelos organizadores.

Tabela 1. Itens e classificação da dificuldade pelos organizadores Bebras

Item	Questão	Classificação Bebras
Item 1	Cerimônia	Fácil
Item 2	Artes de toras	Fácil
Item 3	Castores em fuga	Fácil
Item 4	Tráfego na vila	Fácil
Item 5	Rede à prova	Médio
Item 6	Labirinto espacial	Fácil
Item 7	Pegadas	Médio
Item 8	Saltos	Médio
Item 9	Rede social	Médio
Item 10	Jogo da altura	Difícil
Item 11	Ponto de encontro	Difícil
Item 12	Melhor tradução	Difícil
Item 13	Máquina quebrada	Médio
Item 14	Verdadeiro ou Falso	Difícil
Item 15	Retângulos	Difícil

4.3. Procedimentos

Aplicamos as provas impressas com tempo de duração de 55 minutos. Antes da aplicação, conversamos com os alunos sobre o objetivo do estudo, o tempo de duração da prova e as instruções para resolução das questões. Ao final do tempo transcorrido, as provas foram recolhidas e cada item foi corrigido considerando certo, errado ou sem resposta (NA - *no answer*). O resultado foi armazenado em uma tabela para análise.

Após a correção das provas, calculamos os escores dos alunos, média e desvio padrão de acerto dos itens, bem como a média de acerto por item e a correlação ponto biserial (item-escore total). Em seguida, iniciamos a análise TRI. Utilizamos a *add-in* EIRT versão 2.0.0 para MS Excel que utiliza o estimador modal de Bayes com estimação a *posteriori* (*Expected A Posteriori* - EAP) [Germain et al. 2006]. Nesta análise exploratória via TRI, consideramos que pensamento computacional é o fator dominante responsável pelas respostas dos alunos no instrumento aplicado.

5. Resultados

Nesta seção, mostramos os resultados dos dados coletados na amostra de 126 alunos. A média de acerto dos 15 itens foi de 9.61 questões e desvio padrão de 2.679. Em seguida, prosseguimos para estimar os parâmetros da TRI. A escolha do modelo considerou o método qui-quadrado como teste de ajuste dos dados ao modelo. Assim, estimamos o modelo de 3PL e, como o resultado de todos os itens convergiram ao modelo (p-*valor* > 0.990), prosseguimos com a análise exploratória dos demais parâmetros.

A Tabela 2 apresenta os parâmetros estimados de todos os itens do instrumento: média, desvio padrão (D.P.), correlação ponto biserial (Coor.), discriminação a, dificuldade b e acerto ao acaso c. A coluna "média"mostra a média de acertos de cada item. Assim, observamos que o item com maior taxa de acerto foi o item 3, com 96% de acerto, e o item com menor taxa de acerto foi o item 1, como 23% de acerto.

A correlação ponto biserial representa a correção de acerto ou erro do item e o escore total do instrumento. Observando os coeficientes de correlação na coluna "Coor."da Tabela 2, percebemos que os maiores valores encontrados são do item 8 (0.407), seguido do item 10 (0.323). Mesmo assim, são coeficientes baixos, indicando correlações moderada a fraca. Os demais itens possuem valor menor, sobretudo o item 1, com correlação próxima de zero (0.059). Nas seções seguintes, apresentamos os resultados

Tabela 2. Parâmetros dos Itens

Item	Questão	Média	D.P.	Corr.	*a*	*b*	*c*
Item 1	Cerimônia	0.230	0.421	0.059	0.720	4.477	0.184
Item 2	Artes de toras	0.841	0.365	0.227	0.863	-1.838	0.162
Item 3	Castores em fuga	0.968	0.175	0.259	1.700	-2.557	0.166
Item 4	Tráfego na vila	0.712	0.453	0.304	1.041	-0.672	0.143
Item 5	Rede à prova	0.611	0.487	0.327	1.415	0.012	0.173
Item 6	Labirinto espacial	0.913	0.282	0.306	1.477	-1.882	0.162
Item 7	Pegadas	0.786	0.410	0.257	0.958	-1.187	0.173
Item 8	Saltos	0.864	0.343	0.407	2.375	-1.132	0.145
Item 9	Rede social	0.569	0.495	0.323	0.933	0.174	0.136
Item 10	Jogo da altura	0.431	0.495	0.334	1.364	0.813	0.134
Item 11	Ponto de encontro	0.509	0.500	0.316	0.812	0.823	0.136
Item 12	Melhor tradução	0.630	0.483	0.246	0.670	0.060	0.173
Item 13	Máquina quebrada	0.560	0.496	0.167	0.594	0.464	0.182
Item 14	Verdadeiro ou Falso	0.615	0.487	0.308	0.969	0.164	0.195
Item 15	Retângulos	0.372	0.483	0.227	1.772	1.395	0.202

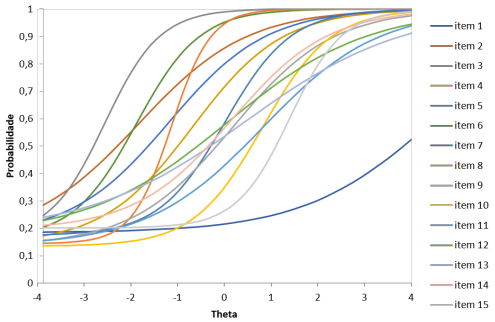

Figura 1. Curva Característica dos Itens

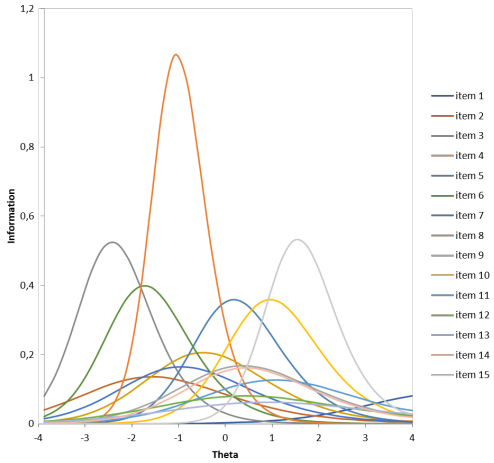

Figura 2. Função de Informação dos Itens

dos parâmetros estimados pela TRI.

5.1. Dificuldade dos Itens

Na TRI, a dificuldade do item é estimada pelo parâmetro *b* na mesma escala de habilidade (Θ) e representa o nível de aptidão necessário para o sujeito ter a probabilidade de 50% de acertar o item. Assim, considerando que o valor de Θ médio é 0, os itens 2, 3, 4, 6, 7, 8 possuem valores de *b* negativos e os itens 5 e 12 possuem valores de *b* próximos a 0 (0.012 e 0.060, respectivamente). Por isso, esses itens podem ser considerados fáceis de acordo com estimação pela TRI. Já os itens com *b* entre 0.1 e 1 podem ser considerados de dificuldade média, como por exemplo, os itens 9, 10, 11, 13 e 14. Já os itens 1 e 15 podem ser considerados difíceis, pois possuem valor de *b* maior que 1. A Figura 1 mostra as CCI dos itens. Nela podemos perceber a curva do item 1 mais à direita (na cor azul escuro), indicando que esse item é o mais difícil do instrumento. Em particular, o item 1 possui valor de $b = 4.477$, indo além do valor padrão máximo esperado para Θ que se costuma analisar (entre -3 e 3). Portanto, esse item pode não ser adequado para avaliação.

Comparando a precisão de dificuldade dos itens fornecida pelos organizadores do Bebras com a mensuração pela TRI, percebemos que houve adequação em sete itens, outros sete itens foram sobrestimados e um item foi subestimado. A previsão da dificuldade

foi a mesma para os itens 2, 3, 4 e 6 apontados como fáceis, itens 9 e 13 com dificuldade média e o item 15 como difícil, pelos organizadores do Bebras e pela TRI. Entretanto, os itens 5, 7, 8, 10, 11, 12, 13 e 15 foram indicados pelos organizadores do Bebras com dificuldade maior do que a estimada pela TRI, ou seja, sobrestimados. E o item 1 apontado como fácil pelos organizadores, foi estimado como difícil.

5.2. Discriminação dos Itens e Função de Informação

A coluna *a* da Tabela 2 mostra os índices de discriminação dos itens. Observamos que o item 13 tem discriminação baixa, mas os itens 1, 2, 4, 7, 9, 11, 12 e 14 possuem discriminação moderada, e os itens 3, 5, 6, 8, 10 e 15 apresentam discriminação alta, segundo [Baker 2001]. Assim, podemos concluir que o instrumento possui itens com discriminação moderada e alta.

Complementando a análise da discriminação, observamos as funções de informação dos itens na Figura 2 no intuito de analisar, de forma preliminar, para qual intervalo de habilidade as questões mais discriminativas apresentam maior informação [Baker 2001]. Nela, podemos perceber que o item 8 (em laranja) é o que apresenta mais informação para sujeitos com baixa habilidade. Os itens 3 (em cinza claro) e 6 (verde) também apresentam discriminação e informação alta para sujeitos com baixa habilidade. Já o item 5 (em azul) apresenta mais informação para sujeitos com habilidade média. Considerando sujeitos com habilidade alta, os itens 10 (em amarelo) e 15 (em cinza claro) fornecem mais informação. Por último, os valores do parâmetro de acerto ao acaso *c* encontram-se abaixo de 40%.

6. Discussão

Nesta seção, discutimos os resultados quanto a consistência interna do instrumento, os índices de dificuldade estimados pelo Bebras e pela TRI e a discriminação dos itens.

Os valores baixos de correlação ponto biserial contribuem para uma baixa consistência interna do instrumento. Baixas correlações item-total indicam que os itens se associam pouco uns com os outros. Assim, calculamos o coeficiente alfa de Cronbach no intuito de estimar a consistência interna do instrumento [Hutz et al. 2015]. Como era esperado, o valor do coeficiente foi 0.648, considerado baixo para garantir a consistência interna de um instrumento. Esse resultado pode ser oriundo da falta de unidimensionalidade dos itens. Neste estudo exploratório, consideramos que o instrumento era unidimensional, ou seja, avaliava apenas um construto: Pensamento Computacional. Entretanto, outros estudos apontam a existência de habilidades de granularidade menor associadas ao PC [Avila et al. 2017, Araujo et al. 2016], bem como a própria organização do Bebras reconhece que uma questão pode ter uma ou mais habilidade(s) associada(s) [Dagiene and Stupuriene 2016]. Portanto, o instrumento precisa ser examinado quanto as possíveis habilidades envolvidas nas questões.

A estimativa da dificuldade das questões fornecida pelos organizadores do Bebras apontou precisão de 46,6% quando comparada à estimativa calculada pela TRI. Uma das causas desse resultado pode ser devido às diferenças culturais, uma vez que a estimativa foi realizada por pesquisadores do Reino Unido e o instrumento foi traduzido e aplicado no Brasil. Mesmo assim, outros estudos já sugerem diferenças entre a estimação teórica da dificuldade e a estimada após a aplicação da prova [Dagiene and Stupuriene 2016,

Araujo et al. 2017]. Além disso, na TRI, os parâmetros de dificuldade possuem valores próximos independente da amostra, divergindo quando é estimada com muito erro, como por exemplo, quando a amostra é composta por sujeitos sem variação nos níveis de habilidade [Hutz et al. 2015]. Em suma, esse resultado reflete também na complexidade envolvida na atividade de estimar a dificuldade de uma questão baseada apenas no conhecimento tácito de elaboradores, sem considerar os resultados empíricos.

Itens apontados na seção anterior com discriminação alta são também os itens que apresentam maior informação ao longo do intervalo de habilidade (itens 8, 3, 6, 5, 10 e 15). Dessa forma, esse conjunto de itens poderiam ajudar a diferenciar sujeitos com habilidades próximas em quase toda escala de habilidade Θ. Por último, considerando o valores de acerto ao caso, o instrumento apresentam valores abaixo de 30% de chute.

6.1. Ameaças à Validade

Como toda pesquisa empírica, este trabalho possui ameaças à validade. O número de sujeitos participantes do estudo não permite generalização dos resultados, entretanto respeitamos o número mínimo de 8 respondentes por item para executar os procedimentos da TRI [Hutz et al. 2015]. A correção das provas aconteceu de forma manual, portanto, para mitigar possíveis erros humanos, realizamos dupla checagem das respostas. As questões foram traduzidas da língua inglesa, e para minimizar viés de tradução, cada questão foi revisada por dois pesquisadores externos.

7. Considerações Finais

Este estudo investigou parâmetros psicométricos de questões de uma competição Bebras para avaliar PC. Assim, empregamos a TRI como método estatístico de análise de itens para estimar os parâmetros. Os resultados gerais apontaram que as questões analisadas apresentam discriminação moderada e alta, e quanto a dificuldade, os itens possuem dificuldade fácil e moderada.

Ainda sobre a dificuldade dos itens, os resultados apontaram que 46,6% (sete de quinze itens) foram sobrestimados, ou seja, foram apontados previamente pelos organizadores como mais difíceis do que foram estimados pela TRI. Apenas outros 46,6% da dificuldade dos itens estimados pelos organizadores tiveram a mesma estimativa pela TRI. Dessa forma, percebemos a importância de utilizar métodos que permitam estimar a dificuldade de itens mais próximas do resultado experienciado quando planejamos avaliar habilidades nos sujeitos.

A pontuação baseada na dificuldade proposta pelo Bebras não se mostrou adequada para avaliar, porque as questões podem ser estimadas de maneira imprecisa quanto a dificuldade e por isso, o aluno pode receber uma pontuação que não condiz com sua aptidão em PC. Todavia, o uso do escore pode também não ser adequado para avaliar PC porque esse tipo de avaliação não consegue diferenciar pessoas que acertam a mesma quantidade de questões [Hutz et al. 2015]. Nesse quesito, a TRI auxilia tanto na estimativa da dificuldade dos itens mais próxima da realidade dos sujeitos, como consegue diferenciar sujeitos que acertam a mesma quantidade de questões.

A principal contribuição deste estudo é explorar a aplicabilidade de técnicas de psicometria para auxiliar na avaliação de itens para mensurar PC. Os resultados apresentados sugerem que o grupo de questões que compõe o instrumento utilizado não se

encontra adequado para avaliar PC, pois possui confiabilidade baixa. Entretanto, o estudo demonstra que dentre os itens do instrumento, existem questões que podem ser exploradas para avaliar PC pois possuem bons níveis de discriminação em todo intervalo de habilidade Θ. Como trabalhos futuros, continuaremos investigando a TRI como método para mensurar PC em outras questões Bebras, bem como planejamos explorar as habilidades do PC usadas para resolver as questões no intuito de examinar a granularidade das habilidades relacionadas ao PC.

Referências

Araujo, A. L. S. O., Andrade, W. L., and Guerrero, D. D. S. (2016). A systematic mapping study on assessing computational thinking abilities. In *Frontiers in education conference (FIE), 2016 IEEE*, pages 1–9. IEEE.

Araujo, A. L. S. O., Santos, J. S., Andrade, W. L., Guerrero, D. D. S., and Dagiene, V. (2017). Exploring computational thinking assessment in introductory programming courses. In *2017 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE.

Avila, C., Cavalheiro, S., Bordini, A., Marques, M., Cardoso, M., and Feijo, G. (2017). Metodologias de avaliação do pensamento computacional: uma revisão sistemática. In *Simpósio Brasileiro de Informática na Educação*, volume 28, page 113.

Baker, F. B. (2001). *The basics of item response theory*. ERIC. Disponível em: https://files.eric.ed.gov/fulltext/ED458219.pdf.

Barcelos, T., Souza, A., Silva, L., Muñoz, R., and Acevedo, R. V. (2017). Mensurando o desenvolvimento do pensamento computacional por meio de mapas auto-organizáveis: um estudo preliminar em uma oficina de jogos digitais. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 932.

Blokhuis, D., Millican, P., Roffey, C., Schrijvers, E., and Sentance, S. (2014). UK Bebras Computational Thinking Challenge. Disponível em : http://www.bebras.uk/uploads/2/1/8/6/21861082/ukbebras2014-answers_1.pdf.

Csizmadia, A., Curzon, P., Dorling, M., Humphreys, S., Ng, T., Selby, C., and Woollard, J. (2015). Computational thinking: A guide for teachers. *Google Scholar*.

Dagiene, V. and Stupuriene, G. (2016). Bebras-a sustainable community building model for the concept based learning of informatics and computational thinking. *Informatics in Education*, 15(1):25.

Germain, S., Valois, P., and Abdous, B. (2006). libirt - item response theory library.

Hutz, C. S., Bandeira, D. R., and Trentini, C. M. (2015). *Psicometria*. Artmed Editora.

Raabe, A., Santana, A. L. M., Ellery, N., and Gonçalves, F. (2017). Um instrumento para diagnóstico do pensamento computacional. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 1172.

Román-González, M., Pérez-González, J.-C., Moreno-León, J., and Robles, G. (2016). Does computational thinking correlate with personality?: the non-cognitive side of computational thinking. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 51–58. ACM.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3):33–35.