Avaliação do Desenvolvimento do Pensamento Crítico no Ensino de Programação para Estudantes do Ensino Técnico e Superior: Uma Proposta de Modelo

Deise Monquelate Arndt[1, 2], Ramon Mayor Martins[1], Jean C. R. Hauck[2]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
88.040-370 – Florianópolis – SC – Brasil

[2]Área de Telecomunicações – Instituto Federal de Santa Catarina (IFSC)
São José, SC, Brasil.

{deise.arndt,ramon.mayor}@ifsc.edu.br,jean.hauck@ufsc.br}

Abstract. *Este artigo propõe o desenvolvimento e a validação de um modelo para avaliar o pensamento crítico no ensino de programação para estudantes do ensino técnico e superior. Utilizando a metodologia de Design Centrado em Evidência (ECD) e as Concepções de Avaliação Baseada em Princípios de Investigação (PADI), o modelo foi elaborado para capturar habilidades como interpretação, análise e avaliação. A Validade de Conteúdo e do Construto foi avaliada por um painel de especialistas, mostrando resultados promissores, como a capacidade do modelo medir com precisão os aspectos pretendidos. Assim, o modelo demonstra um grande potencial para ser uma ferramenta eficaz na educação em computação.*

Resumo. *This paper proposes the development and validation of a model to assess critical thinking in programming education for technical and higher education students. Using the Evidence-Centered Design (ECD) methodology and the Principles of Investigation-Based Assessment (PADI) conceptions, the model was developed to capture skills such as interpretation, analysis, and evaluation. The Content and Construct Validity were evaluated by a panel of experts, showing promising results, such as the model's ability to accurately measure the intended aspects. Thus, the model demonstrates significant potential to be an effective tool in computer education.*

1. Introdução

A computação desempenha um papel fundamental na sociedade contemporânea, impulsionando avanços tecnológicos e moldando o futuro. Nesse contexto, a educação em computação torna-se essencial, contribuindo para o desenvolvimento de habilidades como pensamento computacional, criatividade, resolução de problemas e colaboração [Lin e Chen 2020]. Entre essas habilidades, há também o pensamento crítico que se destaca como uma competência indispensável tanto para a computação quanto para a vida cotidiana dos estudantes [Sari *et al*. 2022; World Economic Forum 2020].

O pensamento crítico envolve um conjunto de habilidades cognitivas e disposições que possibilitam aos indivíduos analisar, interpretar, inferir, avaliar, explicar e autorregular informações de forma lógica e independente [Facione 1990]. Esse conjunto de habilidades é importante para a resolução de problemas, a tomada de decisões fundamentadas e a participação ativa na sociedade [Araújo *et al*., 2024]. Diante de sua relevância, o pensamento crítico tem sido amplamente reconhecido como um objetivo

central da educação, abrangendo desde a educação básica até o ensino superior [UNICEF 2023; OECD 2019].

Diversas iniciativas globais têm buscado fomentar o pensamento crítico na educação. Organizações como a "*The Foundation for Critical Thinking*" e a "*Insight Assessment*" desenvolvem cursos e ferramentas voltadas para a avaliação e aprimoramento dessas habilidades [CriticalThinking.org 2019; Insight Assessment 2023]. No Brasil, o Instituto Ayrton Senna promove práticas pedagógicas baseadas no desenvolvimento humano e na criatividade [Instituto Ayrton Senna 2022]. Além disso, o pensamento crítico é cada vez mais integrado ao ensino de computação, sendo aplicado no aprendizado de programação, robótica e inteligência artificial (IA) [Huang e Qiao 2024; Lee *et al*. 2023]. No âmbito educacional, iniciativas como a "*Computer Science Framework (K-12)*" e o projeto da OCDE "*Fostering and Assessing Creativity and Critical Thinking*" para o ensino superior, visam estruturar currículos que desenvolvam o pensamento crítico por meio da computação [K12CS.org 2016; OECD 2024). Além disso, estratégias pedagógicas baseadas em metodologias ativas, como aprendizagem baseada em problemas e projetos, têm se mostrado eficazes na promoção do pensamento crítico [Mäkiö & Mäkiö 2023; Taylor 2022].

Recentemente, a ascensão das tecnologias de IA, especialmente os Modelos de Grandes Linguagens *(LLMs*), trouxe novos desafios e oportunidades para a educação em computação. Se, por um lado, essas ferramentas ampliam o acesso ao conhecimento, por outro, levantam preocupações quanto à autonomia dos estudantes na construção do pensamento crítico e na avaliação da veracidade das informações [Zawacki-Richter *et al.* 2019; UNICEF 2023]. Dessa forma, o fortalecimento das habilidades de pensamento crítico torna-se ainda mais necessário para capacitar os estudantes na análise criteriosa das informações e na tomada de decisões informadas. Além do ensino, a avaliação do pensamento crítico é um aspecto essencial para a melhoria da educação em computação. Métodos confiáveis e validados são essenciais para medir o desenvolvimento dessas habilidades e fornecer um *feedback* eficaz para estudantes e professores [Paul *et al*. 2023; Moskal e Leydens 2000].

Alguns trabalhos realizaram estudos com avaliações para o pensamento crítico como Catojo e Nunes (2024), Kaur e Chahal (2023), Song *et al*. (2021), Gong *et al*. (2020) e Liu *et al*. (2018). Contudo, a maioria desses autores utilizaram testes padronizados/comerciais, limitando-se a somente autoavaliação pelos estudantes [Arndt *et al*. 2024]. Esses estudos concentram-se em um conjunto limitado de habilidades avaliadas e estão concentradas como iniciativas nos países asiáticos [Arndt *et al.* 2024; Arndt *et al.* 2025]. Além disso, há uma escassez de avaliação direcionada a um público alvo como estudantes do ensino técnico e superior [Arndt *et al.* 2025]. Assim, apesar da reconhecida importância do pensamento crítico, há uma lacuna na literatura sobre como avaliar efetivamente essas habilidades no contexto específico do ensino de programação, para estudantes do nível técnico e superior.

Nesse sentido, este estudo propõe o desenvolvimento e a avaliação inicial de um modelo para avaliar o desenvolvimento do pensamento crítico no ensino de programação para estudantes do nível técnico e superior, assegurando um modelo confiável, válido e alinhado aos currículos existentes. O modelo proposto é

sistematicamente desenvolvido por meio da abordagem *ECD*/*PADI* [Mislevy *et al*. 2003; Seeratan e Mislevy 2008] e avaliado por meio de um painel de especialistas. Os principais resultados desta pesquisa são diretrizes para um modelo de avaliação para pesquisadores em educação em computação e metodologias práticas para professores de computação em exercício.

Este artigo está estruturado da seguinte forma: na seção 2, a metodologia utilizada é apresentada. Na seção 3, a estrutura do modelo de avaliação do desenvolvimento do pensamento crítico é detalhada. Na seção 4 é apresentada a avaliação do modelo proposto. A seção 5 apresenta os resultados das avaliações de Validade de Conteúdo e Construto a partir do painel de especialistas. Por fim, na seção 6 os resultados são discutidos, concluídos e estabelecidos trabalhos futuros.

2. Metodologia de pesquisa

Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Santa Catarina, conforme Parecer nº 7.320.427.

O modelo de avaliação foi desenvolvido e implementado seguindo a abordagem de *Design* Centrado em Evidência (*ECD*) proposta por Mislevy *et al*. (2003) e na aplicação das Concepções de Avaliação Baseada em Princípios de Investigação (*PADI*) propostas por Seeratan e Mislevy (2008) e Riconoscente *et al*. (2005). O *ECD* é um *framework* para o *design* de avaliações educacionais, que visa garantir que as avaliações sejam projetadas de forma rigorosa e eficaz para medir os construtos pretendidos e fornecer evidências necessárias para apoiar as conclusões tiradas dos resultados [Mislevy *et al.* 2003]. Já o *PADI* é uma ferramenta prática que utiliza o *ECD* como base teórica para criar avaliações que sejam rigorosas e eficazes em medir habilidades de inquérito científico. O *PADI* é uma colaboração entre várias instituições (SRI International, Universidade de Maryland, Universidade da Califórnia-Berkeley, Universidade de Michigan e Lawrence Hall of Science) e visa fornecer uma abordagem prática e baseada em teoria para o desenvolvimento de avaliações de qualidade em inquérito científico [Seeratan e Mislevy 2008]. O uso conjunto das duas abordagens serve para estruturar a avaliação de forma sistemática e prática. Ambas abordagens são definidas por camadas (Figura 1).

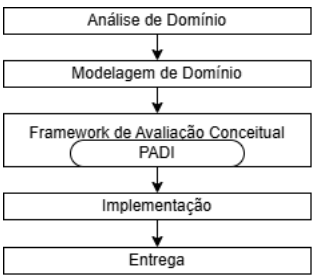

Figura 1. Abordagem metodológica *ECD* e *PADI* (elaborado pelos autores, baseado em Mislevy e Riconscente, 2005)

O *ECD* possui as camadas de "Análise de Domínio", "Modelagem de Domínio", "*Framework* de Avaliação Conceitual", "Implementação" e "Entrega". Já o *PADI* atua na camada de "*Framework* de Avaliação Conceitual" do *ECD*. As etapas do *ECD* e *PADI* são:

Análise de Domínio: Esta etapa reúne informações substantivas sobre o domínio, incluindo conceitos, terminologias, ferramentas e formas de representação [Riconoscente *et al*. 2005]. Nesta etapa são analisadas as habilidades, competências e conhecimentos que se deseja medir. São extraídas as informações sobre como o desenvolvimento do pensamento crítico pode ser avaliado a partir do ensino específico [Mislevy *et al*. 2003].

Modelagem de Domínio (Padrões de evidência): Com base na análise de domínio, são desenvolvidas narrativas que descrevem o domínio de medição e especificam o conteúdo programático. Estas narrativas incorporam o argumento de avaliação, contemplando os *KSAs* (*Knowledge, Skills, and Abilities*), produtos esperados e observações que evidenciam o desenvolvimento das competências almejadas pelos estudantes [Riconoscente *et al*. 2005].

***Framework* de Avaliação Conceitual**: Contém modelos de estudante, evidência e tarefa, com variáveis observáveis e especificações detalhadas [Riconoscente *et al*. 2005]. O Modelo do Estudante considera o conhecimento, as competências e as habilidades utilizadas para extrair evidências sobre a proficiência dos estudantes. O Modelo de Tarefa define as tarefas e recursos utilizados no ensino para obter as evidências de proficiência dos estudantes. E o Modelo de Evidência, inclui especificações de um modelo de avaliação, com variáveis observáveis relacionadas ao desempenho do estudante baseadas nos produtos de trabalho realizados durante as tarefas específicas. Isso cria evidências que demonstram o nível de competência do estudante e um modelo de medição inicial, que sintetiza a pontuação entre os critérios ou categorias do modelo de avaliação.

Implementação: Como produto, é elaborado um modelo de avaliação, que contém o modelo de estudante, evidência e tarefa, com variáveis observáveis e especificações detalhadas.

Entrega: Esta etapa contempla as interações, pontuações e relatórios [Riconoscente *et al.* 2005]. Uma versão inicial do modelo de avaliação é submetida à análise de um painel de especialistas, que revisa o instrumento e fornece *feedback* [Beecham et al. 2005], tendo como instrumento de coleta de dados um questionário estruturado, no qual os especialistas analisam aspectos do conteúdo, como clareza, relevância, abrangência e adequação da linguagem, seguindo as recomendações de Lynn (1986), além de aspectos do construto, incluindo adequação dimensional, coerência da rubrica e progressão dos níveis, seguindo recomendações de Messick (1995).

3. Desenvolvimento do Modelo de Avaliação do Pensamento Crítico no Ensino de Programação

O modelo de avaliação proposto foi desenvolvido para o contexto do ensino de programação, visando avaliar o desenvolvimento do pensamento crítico em estudantes do nível técnico e superior. A estrutura do modelo foi orientada pela metodologia *ECD* e *PADI* em todo desenvolvimento.

Análise de Contexto. Como primeiro passo para o desenvolvimento, foi realizada a caracterização dos estudantes nos diferentes níveis educacionais, considerando aspectos cognitivos, sociais e culturais. O público-alvo são estudantes do ensino técnico, com

idade entre 15 e 17 anos, e do ensino superior, com o Ensino Médio concluído e com idade de 18 anos [MEC 2017; Ideb 2023]. Os conhecimentos e habilidades esperados são orientados pela Base Nacional Comum Curricular (BNCC), diretrizes do Ministério da Educação (MEC) e *U.S. Department of Education*. Além disso, foram identificadas as competências específicas de computação recomendadas pela Sociedade Internacional de Tecnologia na Educação (ISTE) e a Sociedade Brasileira de Computação (SBC). A análise incluiu o perfil e a formação dos professores, avaliando sua preparação para ensinar computação e estimular o pensamento crítico. Por fim, foram consideradas as políticas e diretrizes educacionais que influenciam o ensino de computação e o desenvolvimento do pensamento crítico [MEC 2017; SBC 2021 ACM/IEEE-CS 2023].

Análise de Domínio. Como segunda etapa, para o desenvolvimento do modelo, foram consideradas as habilidades do pensamento crítico no ensino de programação envolvendo a revisão das definições e habilidades associadas [Arndt *et al.* 2024], com ênfase no Relatório Delphi [Facione, 1990]: interpretação, análise, avaliação, inferência, explicação e autorregulação, bem como nas habilidades adicionais [Yeh, 2023]: reconhecimento de premissas, indução, dedução. Além disso, são identificadas suas aplicações no ensino técnico e superior.

Modelagem de Domínio. Foram definidos padrões de evidência para cada habilidade, com exemplos de questões múltipla escolha com escala Likert e questões abertas, representando a aplicação dos componentes narrativos e observáveis. A justificativa para os aspectos avaliados está na resolução de problemas em contextos técnicos e computacionais, promovendo decisões fundamentadas. Seguindo o "Padrão de evidência" *PADI* [Seeratan e Mislevy, 2008; Riconoscente *et al*., 2005], foram avaliadas as "habilidades focais", como as habilidades do pensamento crítico de: interpretação, análise, avaliação, inferência, explicação, autorregulação, reconhecimento de premissas, indução e dedução em tarefas baseadas em programação e resolução de problemas. E como "habilidades complementares", outras habilidades necessárias para os estudantes realizarem as tarefas propostas, incluindo: a leitura e interpretação de texto, e a familiaridade com conceitos básicos de lógica e programação. Além disso, como características fixas, é estabelecido o uso de um cenário pré-definido, questões baseadas em exemplos práticos de programação simples, de caráter geral e universal. Entende-se por esse caráter, aquele que os estudantes de qualquer curso na temática programação, conseguiriam compreender, independente de sintaxe de linguagem de programação.

***Framework* de Avaliação.** Foram desenvolvidos o Modelo do Estudante (com as habilidades como variáveis de avaliação), o Modelo de Evidência (com escalas Likert e rubrica analítica) e o Modelo de Tarefa (com questões para capturar habilidades específicas). Seguindo a abordagem *PADI* [Seeratan e Mislevy, 2008; Riconoscente *et al*., 2005], como observações potenciais, o modelo prevê respostas claras e justificativas para escolhas (nas questões abertas). Do produto de trabalho potencial, o modelo prevê respostas escritas nas questões abertas; resultados em questões de múltipla escolha; e análise de casos descritos em texto. Da rubrica potencial, é utilizado a escala Likert de 5 pontos para questões fechadas; e uma rubrica analítica para questões abertas (variando de 1-Inadequado à 5-Excelente).

As tarefas exemplares, que representam o padrão de evidência, são questões que realizam a imersão em um cenário pré-definido com exemplos práticos, por exemplo: "*Quando seu código apresenta um comportamento inesperado, como você descobre a causa do problema? Explique seu raciocínio.*" e "*Durante o desenvolvimento de um código, eu: […]*".

As referências que sustentam o padrão de evidência são textos acadêmicos sobre computação e pensamento crítico discutidos por professores e especialistas da área.

Entrega. O modelo de avaliação do desenvolvimento do pensamento crítico no ensino de programação para estudantes do ensino técnico e superior está disponível no link: https://tinyurl.com/ycx5tc4n

4. Avaliação do Modelo Proposto

Visando avaliar o modelo em termos de Validade de Conteúdo e de Construto, foi realizado um painel de especialistas, no mês de fevereiro de 2025.

4.1. Painel de Especialistas

O painel de especialistas seguiu a abordagem metodológica apresentada por Beecham *et al*. (2005) e as diretrizes de Haynes *et al.* (1995) e Alexandre e Coluci (2011) para validação de modelos, contemplando aspectos quantitativos e qualitativos. A seleção dos especialistas seguiu as recomendações e critérios de Grant e Davis (1997) e Davis (1992). Dessa forma, foram convidados 37 especialistas, dos quais 16 responderam o questionário. Essa quantidade de respostas obtidas atende às recomendações de Lynn (1986) sobre o número mínimo de especialistas para validação de conteúdo e construto.

O corpo de especialistas foi selecionado atendendo às considerações de Haynes *et al.* (1995) e Lawshe (1975) sobre a necessidade de uma amostra representativa, contemplando as diferentes áreas de *expertise* para uma avaliação abrangente do modelo. Os especialistas foram convidados de acordo com sua formação acadêmica, considerando especialização na área da computação, educação e engenharias. Também foi considerada a experiência profissional, a partir da docência no ensino de programação e/ou experiência com pensamento computacional, pensamento crítico e avaliação educacional.

Do conjunto de especialistas participantes, 9 são da engenharia, 5 da computação e 2 da área da educação. Desses, 13 possuem título de doutor e 3 de mestre. O tempo de experiência profissional no ensino desses participantes variou de 1 ano a 35 anos, com média de aproximadamente 18 anos. Do tempo de experiência com ensino de programação/lógica/algoritmo e pensamento computacional, 31% tinha mais de 5 anos.

4.2. Definição da Avaliação

As avaliações foram definidas para verificar a Validade de Conteúdo e Validade do Construto [Lynn 1986; Messick 1995]. A avaliação da Validade de Conteúdo verifica se o instrumento de avaliação abrange adequadamente todo o domínio (conteúdo) que pretende medir, ou seja, verifica se as questões avaliam realmente as habilidades de pensamento crítico em programação. A validade de conteúdo foca em "como" o instrumento está apresentado, incluindo a avaliação de clareza, relevância, abrangência e adequação da linguagem [Lynn 1986], conforme mostra a Tabela 1.

Tabela 1. Critérios da Avaliação da Validade de Conteúdo

Critério	Elementos de Análise	Breve descrição do critério	Breve explicação do critério	Escala
Clareza	•Avaliação da redação dos itens •Compreensibilidade dos critérios •Objetividade das descrições	•A clareza analisa se as questões estão bem escritas; se são compreensíveis; •Se as questões são objetivas	Ex: "Durante o desenvolvimento de um código, eu: Identifico partes do código que podem ser melhoradas" - A questão está clara? Compreensível?	Likert de 5 pontos (1-Discordo totalmente a 5-Concordo totalmente)
Relevância	•Pertinência dos itens para avaliação •Importância dos critérios •Adequação aos objetivos do instrumento	•Se as questões são importantes para avaliar pensamento crítico •Se têm relação com programação	Ex: A questão sobre "dividir problemas grandes em partes menores" é relevante para avaliar análise no contexto de programação?	
Abrangência	•Cobertura da habilidade de interesse •Representatividade dos aspectos avaliados •Suficiência dos critérios	•Se cobre todas as sub-habilidades necessárias •Se não falta nenhum aspecto importante	Ex: As habilidades (análise, avaliação, inferência, etc.) cobrem todos os aspectos do pensamento crítico em programação?	
Adequação da Linguagem	•Adequação ao público-alvo •Clareza dos termos utilizados •Consistência terminológica	•Se está adequada ao público-alvo (estudantes) •Se usa terminologia apropriada	Ex: Os termos técnicos usados são adequados para estudantes?	

A Avaliação de Validade de Construto verifica se o modelo realmente mede o conceito teórico que pretende medir, as construções teóricas do pensamento crítico em programação. Os critérios analisados são baseados nos aspectos substantivos e estruturais segundo Messick (1995), conforme mostra a Tabela 2.

Tabela 2. Critérios da Avaliação da Validade de Construto

Critério	Alinhamento com Messick (1995)	Elementos de Análise	Breve descrição do critério	Breve explicação do critério	Escala
Adequação das habilidades	Aspecto substantivo	•Alinhamento entre itens e habilidades •Representatividade dos construtos •Organização lógica dos elementos	•Se as questões se alinham com a teoria do pensamento crítico •Se representam corretamente cada habilidade	Ex: As questões de "análise" realmente medem análise segundo à teoria? Ou seja, ao consenso Delphi Report (Facione, 1990)?	Likert de 5 pontos (1-Discordo totalmente a 5-Concordo totalmente)
Coerência das Rubricas	Aspecto estrutural	•Consistência interna dos níveis •Clareza das distinções entre níveis •Adequação dos critérios de progressão	•Se os níveis da rubrica são consistentes •Se há distinção clara entre os níveis	Ex: A progressão de "inadequado" para "excelente" na rubrica faz sentido?	
Progressão dos Níveis		•Gradação adequada entre níveis •Distinção clara entre categorias •Equilíbrio das escalas	•Se há gradação adequada entre os níveis •Se as categorias são distintas	Ex: Há diferença clara entre "razoável" e "satisfatório" na rubrica?	

5. Resultados da Avaliação do Modelo Proposto

Nesta seção, são apresentados os resultados da avaliação do modelo proposto por meio do painel de especialistas. Todos componentes para a avaliação do modelo estão disponíveis em: https://tinyurl.com/9c6vjjk3

5.1. Quais são as evidências da Validade de Conteúdo

A Validade de Conteúdo foi avaliada para garantir que o modelo utilizado consiga abranger adequadamente todo o domínio de habilidades de pensamento crítico em programação que se pretende medir. A avaliação da Validade de Conteúdo focou nos critérios de clareza, relevância, abrangência e adequação da linguagem. Para isso, foram empregados dois índices principais: o Coeficiente de Validade de Conteúdo (CVR) e o Índice de Validade de Conteúdo (CVI).

O CVR foi calculado segundo o método proposto por Lawshe (1975) e o CVI foi calculado seguindo o método de Lynn (1986) e Alexandre e Coluci (2011). Os resultados do CVR e CVI para os itens avaliados estão apresentados na Tabela 3. Além disso, foi medida a porcentagem de concordância entre os observadores. Esses métodos são complementares e auxiliam na análise da Validade de Conteúdo, fornecendo uma visão geral da consistência na avaliação e concordância entre os especialistas.

Tabela 3. Resultados da Avaliação da Validade de Conteúdo

Critério de Avaliação	CVR	CVI	Porcentagem de Concordância
	■ Aprovado ≥ 0,5 ▲ Reprovado < 0,5 (Lawshe, 1975)	■ Aprovado ≥ 0,78 ▲ Reprovado < 0,78 (Alexandre e Colucci, 2011)	
Clareza - Redigidas de forma clara	0,8750 ■	0,9375 ■	93,75%
Clareza - Compreensíveis	1,0000 ■	1,0000 ■	100,00%
Clareza - Linguagem objetiva	1,0000 ■	1,0000 ■	100,00%
Relevância - Pertinentes para o objetivo	0,8750 ■	0,9375 ■	93,75%
Relevância - Relação com programação	1,0000 ■	1,0000 ■	100,00%
Relevância - Aspectos do pensamento crítico	0,5000 ■	0,7500 ▲	75,00%
Abrangência - Cobre o domínio avaliado	0,5000 ■	0,7500 ▲	75,00%
Abrangência - Representa habilidades do pensamento crítico	0,6250 ■	0,8125 ■	81,25%
Abrangência - Critérios suficientes	0,5000 ■	0,7500 ▲	75,00%
Adequação da Linguagem - Adequada ao público-alvo	0,6250 ■	0,8125 ■	81,25%
Adequação da Linguagem - Termos consistentes	0,7500 ■	0,8750 ■	87,50%
Adequação da Linguagem - Terminologia técnica apropriada	0,8750 ■	0,9375 ■	93,75%

Em geral, os resultados mostram que o instrumento foi bem avaliado em termos de clareza e relevância, com altos índices de aprovação. No entanto, alguns aspectos da abrangência e da relevância específica do pensamento crítico em programação apresentaram resultados mais baixos, indicando potenciais áreas para melhoria. Todavia, todos os resultados foram considerados válidos conforme o CVR, passando ou igualando o valor crítico de 0,5 da interpretação de Lawshe (1975). O CVI teve quase todos os critérios "aprovados", passando o valor crítico de 0,78 da interpretação de Alexandre e Colucci (2011). Os critérios "reprovados", por sua vez, estão bem próximos do valor crítico, com valor igual a 0,75.

5.2. Quais são as evidências da Validade do Construto

A Validade do Construto foi avaliada visando garantir que o instrumento realmente mede o conceito teórico que se pretende medir, ou seja, verificar se as questões capturam as construções teóricas do pensamento crítico no ensino de programação. Para isso, foram utilizadas várias métricas que avaliam a concordância entre os avaliadores e a consistência das avaliações (Tabela 4).

O coeficiente Fleiss Kappa foi calculado para quantificar a concordância entre os especialistas na avaliação dos construtos. O valor obtido foi de (κ = -0,1181, p-valor 0,0078), o que indica um coeficiente de baixa concordância entre os especialistas. Contudo, esse resultado sugere o conhecido na estatística por "paradoxo de Kappa" [Derksen *et al.*, 2024], onde a alta concordância nas respostas pode levar a um coeficiente baixo devido à falta de variabilidade nas respostas dos especialistas.

Assim, para investigar com mais profundidade a Validade do Construto, a Correlação Intraclasse (ICC) foi utilizada para medir a consistência das avaliações entre os especialistas. Nessa avaliação, os resultados mostram que a consistência das avaliações individuais foi "boa", com ICCs variando entre 0,74 e 0,75. Quando considerados como grupo de avaliadores, a consistência foi "excelente", com ICCs de 0,96 para todos os tipos de avaliadores. Isso indica que o consenso entre os especialistas é alto.

De forma complementar, o Coeficiente de Concordância de Kendall [Kendall 1948] foi calculado para medir a força da concordância entre os avaliadores. O valor obtido foi de 0,6057, o que indica uma concordância moderada a alta entre os especialistas. Esse resultado sugere que há um grau significativo de acordo entre os avaliadores sobre a relação dos itens com os construtos teóricos. Além disso, foi

calculada a porcentagem de concordância bruta entre os especialistas, com o valor de 78.30%. Embora não haja um critério fixo para essa medida, valores acima de 75% são geralmente considerados indicativos de boa concordância [Graham *et al*., 2014].

Tabela 4. Resultados da Avaliação da Validade de Construto

Métrica	Valor	Interpretação	Referência de interpretação
Fleiss Kappa	κ= -0,1181, p-valor 0,0078, Z = -2,661	Resultado típico do Paradoxo de Kappa	Landis e Koch (1977)
ICC Individual (Média)	0,74-0,75	Consistência boa	Cicchetti e Sparrow (1981)
ICC Grupos (Média)	0,96	Consistência excelente	Cicchetti e Sparrow (1981)
Coeficiente de Concordância de Kendall	W = 0,6057	Concordância moderada-alta	SPSS (2025)
Concordância Bruta (%)	78,30%	Concordância boa	Graham *et al*. (2014)

A investigação aprofundada apontou resultados que sugerem haver concordância substancial entre os especialistas sobre a relação dos itens com os construtos teóricos e o modelo. Essa investigação apontou boa consistência geral e concordância entre os avaliadores.

5.3. Quais são os achados qualitativos da validade do construto e conteúdo

Os avaliadores destacaram vários pontos fortes do instrumento. Eles elogiaram a abrangência do conteúdo, a clareza e objetividade das questões, e a estrutura organizada que combina questões de múltipla escolha e abertas. A inclusão de uma rubrica detalhada foi vista como positiva para garantir a objetividade na avaliação. Além disso, o instrumento foi considerado relevante para avaliar o pensamento crítico em programação, abordando habilidades como análise, avaliação e inferência.

Entre os pontos a melhorar, os avaliadores sugeriram revisar a redação de algumas questões para evitar ambiguidades e incluir exemplos de respostas para cada nível da rubrica. Também foi mencionada a necessidade de uma introdução mais clara sobre o propósito e o público-alvo do instrumento. Alguns avaliadores sugeriram que a escala Likert poderia ser ajustada e que a inclusão de desafios práticos poderia enriquecer a avaliação.

Outras considerações incluíram a importância de realizar estudos de validação e confiabilidade adicionais após a aplicação com o público alvo, adaptar o questionário para diferentes níveis de experiência em programação e considerar uma versão em inglês para uma aplicação mais ampla.

5.4. Ameaças à validade

Para garantir a robustez metodológica deste estudo, foi implementado um conjunto de estratégias visando reduzir potenciais ameaças à validade. O desenvolvimento do modelo avaliativo seguiu rigorosamente os princípios do *Design* Baseado em Evidências (*ECD*) [Mislevy *et al*. 2003; Mislevy e Riconoscente, 2005], fundamentado em uma detalhada análise contextual e posterior modelagem. A validação preliminar foi conduzida por meio de um painel de especialistas. Para isso, a fim de mitigar a ameaça relacionada à diversidade e tamanho de amostra, a seleção dos avaliadores priorizou a diversidade de *expertise*, incluindo especialistas segundo a formação acadêmica e a experiência profissional, seguindo as recomendações de Grant e Davis (1997) e Davis (1992). O número de 16 especialistas participantes atende aos critérios estabelecidos por Haynes *et al*. (1995), Lynn (1986) e Lawshe (1975) para obtenção de resultados

preliminares confiáveis. Para mitigar as ameaças relacionadas à análise de dados, a avaliação estatística segue os critérios de Validade de Construto de Landis e Koch (1977), Cicchetti e Sparrow (1981), Kendall (1948) e os critérios de Validade de Conteúdo de Lawshe (1975) e Alexandre e Colucci (2011). Para mitigar possíveis vieses decorrentes da subjetividade nas avaliações, foi aplicado o método quantitativo de Validação de Conteúdo desenvolvido por Lawshe (1975), que estabelece um procedimento sistemático baseado no cálculo do Coeficiente de Validade de Conteúdo (CVR) e o Índice de Validade de Conteúdo (CVI), seguindo o método de Lynn (1986). Reconhecendo o caráter inicial do modelo de avaliação, enfatiza-se a necessidade de estudos posteriores em maior escala para corroborar os achados preliminares e explorar aspectos não contemplados nesta primeira fase de validação.

6. Discussão e Conclusão

Este artigo propõe o desenvolvimento e a validação inicial de um modelo para avaliar o pensamento crítico no ensino de programação para estudantes do nível técnico e superior. O modelo de avaliação foi elaborado seguindo a abordagem de *Design* Centrado em Evidência (*ECD*) e as Concepções de Avaliação Baseada em Princípios de Investigação (*PADI*), garantindo uma abordagem sistemática.

A avaliação da Validade de Conteúdo e do Construto do modelo mostrou resultados promissores. A Validade de Conteúdo foi amplamente aprovada, com altos índices de clareza, relevância e adequação da linguagem, indicando que o modelo é bem estruturado para capturar as habilidades de pensamento crítico em programação. Essa avaliação mostrou pontos para melhoria em relevância e abrangência. Da Validade do Construto a investigação aprofundada aponta haver concordância entre os especialistas sobre os construtos teóricos, apontando consistência excelente entre o grupo de especialistas e concordância moderada a alta entre avaliadores individualmente. Esses resultados sugerem que o modelo tem potencial para ser uma ferramenta eficaz na avaliação do pensamento crítico.

Qualitativamente, os avaliadores destacaram a abrangência do conteúdo, clareza e objetividade das questões, e a estrutura organizada do modelo como pontos fortes. O modelo foi considerado relevante para avaliar habilidades de pensamento crítico em programação, como análise e avaliação. Sugestões de melhoria incluíram revisar questões para evitar ambiguidades e incluir exemplos de respostas. Além disso, foi recomendada a realização de estudos adicionais de validação e a adaptação para diferentes níveis de experiência em programação.

Esses resultados indicam que o modelo, em sua forma atual, é capaz de ser aplicado para a avaliação do desenvolvimento do pensamento crítico em estudantes do ensino técnico e superior. Embora possíveis ajustes futuros possam fortalecer ainda mais sua capacidade de medir essas habilidades, o modelo proposto pode contribuir para a formação de profissionais mais capacitados e críticos na área de computação. Trabalhos futuros incluem a aplicação do modelo de avaliação por meio de estudos de caso e a sua posterior avaliação da qualidade em termos de validade e confiabilidade.

Agradecimentos

A todos que contribuíram com o painel de especialistas.

Referências

ACM/IEEE-CS Joint Task Force on Computing Curricula (2023) Computer Science Curricula 2023. Technical report. ACM Press and IEEE Computer Society Press. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3664191 [Acesso em: 12 março 2025].

Alexandre, N. M. C. & Coluci, M. Z. O. (2011) "Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas", Ciência & Saúde Coletiva, 16(7), pp. 3061-3068. [GS Search].

Araújo, L., Pessoa, M., & Pires, F. (2024b). Investigando a relação entre pensamento computacional e narrativas digitais e não digitais. In Anais do 35º Simpósio Brasileiro de Informática na Educação (SBIE) (pp. 3294-3303). Porto Alegre, RS: Sociedade Brasileira de Computação. https://doi.org/10.5753/sbie.2024.244972. [GS Search].

Arndt, D. M., Martins, R. M., & Hauck, J. C. R. (2024). Critical thinking assessment in K-12 computing education: A systematic mapping. Informatics in Education. https://doi.org/10.15388/infedu.2025.02

Arndt, D. M., Martins, R. M., & Hauck, J. C. R. (2024). Avaliação de Habilidades do Pensamento Crítico no Ensino Técnico e Superior no Ensino de Computação: Um Mapeamento Sistemático. Revista Brasileira de Informática na Educação (RBIE) (aceito).

Beecham, S., Hall, T., Britton, C., Cottee, M. & Rainer, A. (2005) "Using an expert panel to validate a requirements process improvement model", Journal of Systems and Software, 76(3), pp. 251-275. [GS Search].

Catojo, A. R. S., & Nunes, M. A. S. N. (2024). O pensamento computacional para o desenvolvimento de aprendizagens de leitura e pensamento críticos no ensino fundamental: Um mapeamento sistemático da literatura. Revista Brasileira de Informática na Educação, 32(1), 135-156. https://doi.org/10.1016/j.tsc.2017.05.005. [GS Search]

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. American Journal of Mental Deficiency, 86(2), 127-137. [GS Search]

Critical Thinking (2019) "Critical thinking testing and assessment". Foundation for Critical Thinking. Disponível em: https://www.criticalthinking.org/pages/critical-thinking-testing-and-assessment/594 [Acesso em: 12 março 2025].

Davis, L. L. (1992) "Instrument review: Getting the most from a panel of experts", Applied Nursing Research, 5(4), pp. 194-197. [GS Search].

Derksen, B. M., Bruinsma, W., Goslings, J.C. & Schep, N. W.L. (2024). The Kappa Paradox Explained. The Journal of Hand Surgery, 49(5), pp. 482-485. https://doi.org/10.1016/j.jhsa.2024.01.006. [GS Search].

Facione, P. A., & Facione, N. C. (1990). Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. ERIC, Institute of Education Sciences. [GS Search]

Gong, D., Yang, H. H., & Cai, J. (2020). Exploring the key influencing factors on college students' computational thinking skills through flipped-classroom instruction. International Journal of Educational Technology in Higher Education, 17(19). https://doi.org/10.1186/s41239-020-00196-0. [GS Search].

Graham, M. & Milanowski, A. & Westat, J. (2014). Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings. [GS Search].

Grant, J. S. & Davis, L. L. (1997) "Selection and use of content experts for instrument development", Research in Nursing & Health, 20(3), pp. 269-274. [GS Search].

Haynes, S. N., Richard, D. C. S. & Kubany, E. S. (1995) "Content validity in psychological assessment: A functional approach to concepts and methods", Psychological Assessment, 7(3), pp. 238-247. [GS Search].

Huang, X. & Qiao, C. (2024) "Enhancing computational thinking skills through artificial intelligence education at a STEAM high school", Science & Education, 33. https://doi.org/10.1007/s11191-022-00392-6. [Acesso em: 12 março 2025]. [GS Search].

Insight Assessment (2023) "Critical thinking assessments for higher education". Disponível em: https://www.insightassessment.com/article/critical-thinking-assessments-for-higher-education. [Acesso em: 12 março 2025].

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2023) "Índice de Desenvolvimento da Educação Básica: Ideb 2023", Disponível em: https://download.inep.gov.br/ideb/apresentacao_ideb_2023.pdf. [Acesso em: 12 março 2025]

Instituto Ayrton Senna, s.d. Criatividade e pensamento crítico. Disponível em: https://institutoayrtonsenna.org.br/o-que-defendemos/criatividade-e-pensamento-critico/ [Acesso em: 12 março 2025].

Kaur, A., & Chahal, K. K. (2023). Exploring personality and learning motivation influences on students' computational thinking skills in introductory programming courses. Journal of Science Education and Technology, 32(5), pp. 778–792. https://doi.org/10.1007/s10956-023-10052-1. [GS Search].

K–12 Computer Science Framework (2023) "K–12 Computer Science Framework", Disponível em: http://www.k12cs.org [Acesso em: 12 março 2025].

Kendall, M.G. (1948). Rank correlation methods. Griffin. [GS Search].

Landis, J. R. & Koch, G. G. (1977) "The measurement of observer agreement for categorical data", Biometrics, 33(1), pp. 159-174. [GS Search].

Lawshe, C. H. (1975) "A quantitative approach to content validity", Personnel Psychology, 28(4), pp. 563-575. [GS Search].

Lee, S., Choi, D., Lee, M., Choi, J. & Lee, S. (2023) “Fostering youth's critical thinking competency about AI through exhibition”, in CHI Conference on Human Factors in Computing Systems, Hamburg, Germany. [GS Search].

Lin, P.-H. & Chen, S.-Y. (2020) “Design and Evaluation of a Deep Learning Recommendation Based Augmented Reality System for Teaching Programming and Computational Thinking”, In: IEEE Access, vol. 8, pp. 45689-45699, doi: 10.1109/ACCESS.2020.2977679. [GS Search].

Liu, J. L., McBride, R. E., Xiang, P., & Scarmardo-Rhodes, M. (2018). Physical education pre-service teachers’ understanding, application, and development of critical thinking. Quest, 70(1), pp. 23-38. https://doi.org/10.1080/00336297.2017.1325205. [GS Search].

Lynn, M. (1986) “Determination and Quantification of Content Validity Index”, Nursing Research, 35, pp. 382-386. https://doi.org/10.1097/00006199-198611000-00017. [GS Search].

Mäkiö, E. & Mäkiö, J. (2023) "The task-based approach to teaching critical thinking for computer science students", Education Sciences, 13(7), pp. 742. https://doi.org/10.3390/educsci13070742. [GS Search].

Messick, S. (1995) “Validity of psychological assessment”, American Psychologist, 50(9), pp. 741-749. [GS Search].

Ministério da Educação (MEC) (2017) "Base Nacional Comum Curricular", Disponível em: http://basenacionalcomum.mec.gov.br.

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003) “A Brief Introduction to Evidence-Centered Design”, ETS Research Report Series, i–29. [GS Search].

Moskal, B. M.; Leydens, J. A. (2000) Scoring rubric development: Validity and reliability. Practical Assessment, Research & Evaluation, v. 7, n. 10. Disponível em: http://PAREonline.net/getvn.asp?v=7&n=10. [Acesso em: 12 março 2025]. [GS Search].

OECD. (2019) OECD Learning Compass 2030: “A series of concept notes. OECD Learning Compass 2030 Concept Note Series”. Disponível em: https://www.oecd.org/education/2030-project/teaching-and-learning/learning/learning-compass-2030/OECD_Learning_Compass_2030_Concept_Note_Series.pdf. [Acesso em: 12 março 2025].

OECD (2024) “Teaching, Learning and Assessing Creative and Critical Thinking Skills”, Disponível em: https://www.oecd.org/education/ceri/assessingprogressionincreativeandcriticalthinkingskillsineducation.htm. [Acesso em: 12 março 2025].

Paul, J. A.; Sinha, M.; Cochran, J. D. (2023) “Instruments to assess students critical thinking—A qualitative approach. Decision Sciences Journal of Innovative Education, v. 21, n. 3. [GS Search].

Riconcente, M. M., Mislevy. J. & Hamel, L. (2005). “An introduction to PADI task templates, Menlo Park: SRI International. (PADI Technical Report 3)”. Disponível

em: https://padi.sri.com/downloads/TR3_Templates.pdf. [Acesso em: 10 janeiro 2025].

Sari, M.K., Sudiyanto, S. & Kurniawan, S.B. (2022) "Critical thinking skills profile of fourth-grade elementary school students in science learning", Proceedings of the 5th International Conference on Learning Innovation and Quality Education, Surakarta, Indonesia. [GS Search].

Seeratan, K. L. & Mislevy, R. J. (2008) "Design patterns for assessing internal knowledge representations", SRI International, Menlo Park. (PADI Technical Report 22). Disponível em: https://padi.sri.com/downloads/TR22_DPForAssessInternalKnowRep.pdf. [Acesso em: 10 janeiro 2025].

Sociedade Brasileira de Computação- SBC (2021) "Educação superior em computação: Estatísticas – 2021". Available at: https://www.sbc.org.br/documentos-da-sbc?task=download.send&id=1461&catid=133&m=0. [Acesso em: 12 março 2025].

Song, D., Hong, H., & Oh, E. Y. (2021). Applying computational analysis of novice learners' computer programming patterns to reveal self-regulated learning, computational thinking, and learning performance. Computers in Human Behavior, 120, 106746. https://doi.org/10.1016/j.chb.2021.106746. [GS Search].

SPSS (2025) Kendall's W Test in SPSS Disponível em: https://spssanalysis.com/kendalls-w-test-in-spss/#:~:text=Kendall's%20Concordance%20Coefficient%20W%20quantifies,the%20reliability%20of%20subjective%20evaluations. [Acesso em: 12 março 2025].

Taylor, W. (2022) "Promoting critical thinking through classroom discussion", in FUIKS, C. L. and CLARK, L. (Eds.) Teaching and learning in honors. [GS Search].

UNICEF. Artificial intelligence chatbots. (2023). Disponível em: https://www.unicef.org/eap/blog/artificial-intelligence-chatbots. [Acesso em: 25 fevereiro 2025].

World Economic Forum (2020) "The Future of Jobs Report 2020". Available at: https://www.weforum.org/publications/the-future-of-jobs-report-2020/ [Acesso em: 12 março 2025].

Yeh, Y. C. (2003) Critical thinking test-level I (CTT-I). Taipei: Psychological Publishing.

Zawacki-Richter, O., Marín, V. I., Bond, M. and Gouverneur, F. (2019) "Systematic review of research on artificial intelligence applications in higher education—where are the educators?", International Journal of Educational Technology in Higher Education, 16(39). https://doi.org/10.1186/s41239-019-0172-8. [GS Search].