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

Deise Monquelate Arndt[12], Ramon Mayor Martins[2], Jean Carlo R. Hauck[1]

[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. *Critical thinking is essential in programming education, yet validated assessment instruments are scarce. This paper reports results from a pilot of ProgCTQ, based on Evidence-Centered Design, with open and Likert-scale items, applied to 29 undergraduate students. ProgCTQ showed adequate reliability ($\alpha = 0.813$), but low convergent validity ($\bar{r} = 0.121$) and fragmented factorial structure (21.2% explained variance). Only 55.6% of items had satisfactory construct validity. Despite limitations, "Interpretation" and "Self-regulation" dimensions performed well, indicating the instrument's potential for contextualized critical thinking assessment in computing.*

Resumo. *O pensamento crítico é fundamental no ensino de programação, mas faltam instrumentos validados para sua avaliação nesse contexto. Este artigo apresenta resultados do piloto do ProgCTQ, instrumento baseado em Evidence-Centered Design, com questões abertas e itens Likert, aplicado a 29 estudantes do ensino superior. O ProgCTQ mostrou confiabilidade geral adequada ($\alpha = 0.813$), mas validade convergente baixa ($\bar{r} = 0.121$) e estrutura fatorial fragmentada (21,2% da variância explicada). Apenas 55,6% dos itens apresentaram validade de construto satisfatória. Apesar das limitações, dimensões como "Interpretação" e "Autorregulação" tiveram bom desempenho, indicando potencial do instrumento para avaliação contextualizada do pensamento crítico em computação.*

1. Introdução

A educação do século XXI tem sido marcada por rápidas transformações tecnológicas, exigindo que os sistemas educacionais desenvolvam além dos conhecimentos técnicos, competências cognitivas complexas, como o pensamento crítico (Silva *et al.*, 2023). Nesse cenário, o ensino de programação no contexto da computação emerge como um campo promissor para o desenvolvimento de habilidades de raciocínio lógico, resolução de problemas e tomada de decisões fundamentadas (Lin; Chen, 2020).

O pensamento crítico, entendido como um processo intencional, autorregulado e reflexivo que envolve análise, interpretação, explicação e avaliação de informações (Facione, 1990), é fundamental para a compreensão de algoritmos, depuração de

DOI: 10.5753/sbie.2025.12053

códigos e construção de soluções criativas em computação (Durak, 2020). Diversas iniciativas internacionais, como OCDE/PISA e UNESCO, reconhecem o pensamento crítico como competência-chave para a formação de estudantes capazes de enfrentar desafios complexos (OCDE, 2019; UNESCO, 2017).

Apesar desses avanços, persiste uma lacuna significativa quanto à existência de instrumentos validados para avaliar o pensamento crítico no ensino de programação, especialmente no contexto técnico e superior (Arndt *et al.*; Pontual-Falcão; França, 2025). A ausência de avaliações adequadas limita tanto a eficácia das estratégias pedagógicas quanto a compreensão do real progresso dos estudantes nessa competência.

Para responder a essa demanda, este estudo apresenta resultados do piloto do ProgCTQ, instrumento desenvolvido com base nos *frameworks Evidence-Centered Design* (*ECD*) e *Principled Assessment Design for Inquiry* (*PADI*), que orientam a elaboração de avaliações alinhadas a objetivos educacionais e fundamentadas em evidências (Seeratan; Mislevy, 2008). O instrumento foi previamente validado por especialistas, cujo retorno contribuiu para seu aprimoramento teórico e prático (Arndt *et al.*, 2025b). O estudo piloto visa analisar a adequação do instrumento ao contexto educacional, buscando evidências iniciais de confiabilidade e validade.

Este artigo está estruturado da seguinte forma: seção 2 apresenta a metodologia; seção 3, a aplicação e coleta de dados; seção 4, os resultados; e seção 5, a discussão, conclusões e trabalhos futuros.

2. Metodologia de Pesquisa

Esta pesquisa caracteriza-se como um estudo piloto, entendido como uma investigação preliminar de pequena escala, cujo objetivo é testar a viabilidade, os métodos e os procedimentos de uma pesquisa mais ampla (In, 2017).

A pesquisa adota uma abordagem de pesquisa multimétodo, utilizando dados qualitativos e quantitativos (Saunders *et al.*, 2019) a fim de capturar, com maior precisão, a complexidade do pensamento crítico, no contexto do ensino de programação.

2.1 Instrumento de Avaliação do Pensamento Crítico

O instrumento de avaliação aplicado (Fig. 1) é composto por dois tipos de itens: (i) questões abertas que exigem argumentação e justificativas, e (ii) itens em escala Likert, voltados à autoavaliação de processos cognitivos. Ambos foram elaborados com base nas habilidades de pensamento crítico propostas por Facione (1990), como análise, inferência, explicação, avaliação e autorregulação e nas habilidades adicionais, aplicáveis ao ensino, como reconhecimento de premissas, indução e dedução, propostas por Yeh (2003). As questões abertas foram analisadas a partir de uma rubrica descritiva, construída conforme o modelo *ECD*.

Todos componentes para a avaliação do instrumento estão disponíveis em: https://acesse.one/co6t1

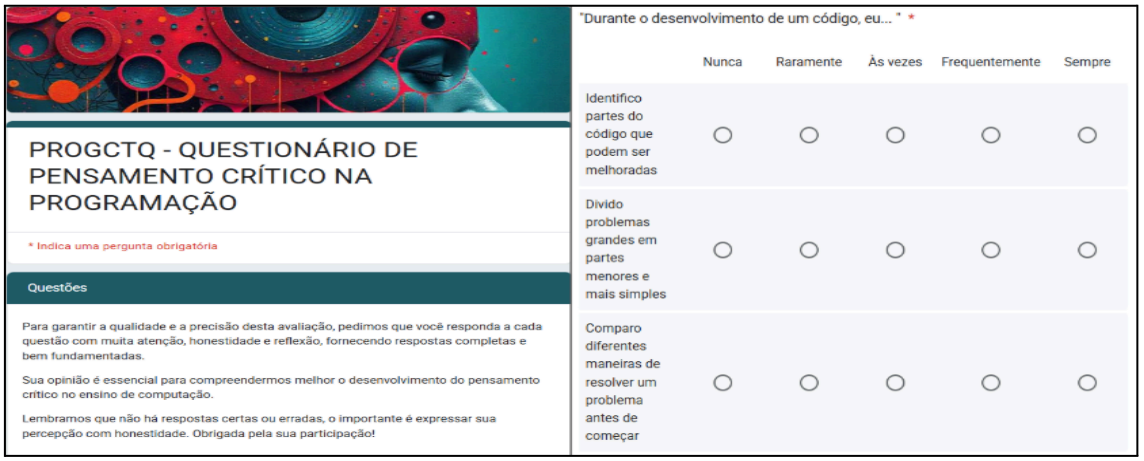

Figura 1. Amostra do Instrumento ProgCTQ

2.2 Participantes

A aplicação piloto foi realizada com 29 estudantes do curso superior em Análise e Desenvolvimento de Sistemas (ADS) de um Instituto Federal. Os participantes foram selecionados por conveniência, a partir da matrícula em disciplinas introdutórias de programação. A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa com Seres Humanos da UFSC (Parecer nº 7.596.663), e contou com termo de autorização institucional da unidade de ensino.

3. Aplicação e Coleta de Dados

A aplicação piloto teve como objetivo analisar a adequação do instrumento de avaliação do pensamento crítico ao contexto do ensino de programação, buscando evidências iniciais de confiabilidade e validade. Esses dados são fundamentais para orientar ajustes metodológicos e fortalecer a estrutura do instrumento antes de sua aplicação em larga escala.

A aplicação ocorreu presencialmente, durante o período letivo regular, em sala de aula. Os estudantes participaram de forma voluntária, mediante assinatura do Termo de Consentimento Livre e Esclarecido (TCLE), conforme as diretrizes éticas. Antes do início, foi apresentada uma breve explicação sobre os objetivos do estudo, reforçando o caráter voluntário e a ausência de impacto nas avaliações das disciplinas.

O instrumento foi disponibilizado por meio de formulário digital (*Google Forms*), com tempo médio de resposta de 30 minutos. Após a coleta, os dados foram organizados para análise quantitativa (itens em escala Likert) e qualitativa (respostas abertas avaliadas por rubrica).

4. Resultados e Análise dos Dados

Os resultados obtidos fornecem evidências iniciais sobre as propriedades psicométricas do instrumento, abrangendo confiabilidade, validade convergente, validade de construto e estrutura fatorial. As análises foram conduzidas utilizando o software R, com foco na avaliação da adequação do instrumento para mensuração do pensamento crítico no contexto da programação.

4.1. Quais são as evidências de confiabilidade

A confiabilidade do instrumento foi avaliada através do coeficiente alfa de Cronbach (Cronbach, 1951), aplicado tanto ao instrumento completo quanto às dimensões individuais (Figura 2). O instrumento geral apresentou α = 0.813, indicando boa consistência interna segundo os critérios estabelecidos por Gliem e Gliem (2003), onde valores ≥ 0.80 são considerados bons para pesquisa.

A análise por dimensões revelou variabilidade significativa nos índices de confiabilidade. A dimensão "Interpretação" obteve o melhor desempenho (α = 0.727), seguida por "Autorregulação" (α = 0.706), ambas consideradas aceitáveis. A dimensão "Dedução" apresentou α = 0.628, situando-se abaixo do recomendado, mas ainda dentro de limites exploratórios para estudos piloto (Nunnally; Bernstein, 1994).

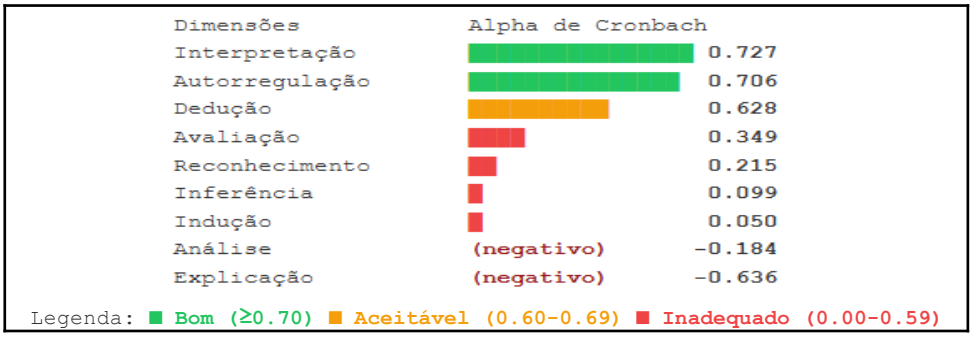

Figura 2. Confiabilidade por dimensão do ProgCTQ (α de Cronbach)

Quatro dimensões apresentaram índices inadequados: "Avaliação" (α = 0.349), "Reconhecimento de Premissas" (α = 0.215), "Inferência" (α = 0.099) e "Indução" (α = 0.050). Duas dimensões apresentaram alfas negativos: "Análise" (α = -0.184) e "Explicação" (α = -0.636), indicando correlações negativas entre os itens, o que sugere necessidade de reformulação.

4.2. Quais são as evidências da validade convergente?

A validade convergente foi avaliada por meio da correlação de Pearson (Cohen, 1988) entre as notas das questões abertas (avaliadas por rubrica) e as médias das questões fechadas (escala Likert) em cada dimensão do instrumento. Os resultados revelaram padrões distintos por dimensão. Correlações moderadas ocorreram em "Avaliação" (r=0,437) e "Reconhecimento de Premissas" (r=0,327), enquanto "Análise" (r=0,273) e "Explicação" (r=0,200) apresentaram valores baixos (Fig. 3). Cinco dimensões apresentaram correlações próximas de zero ou negativas: "Dedução" (r = 0.022), "Inferência" (r = -0.010), "Indução" (r = -0.005), "Autorregulação" (r = -0.029) e "Interpretação" (r = -0.124).

Os resultados indicam que as modalidades avaliativas capturam aspectos distintos do pensamento crítico, sugerindo necessidade de revisão no alinhamento entre critérios e itens.

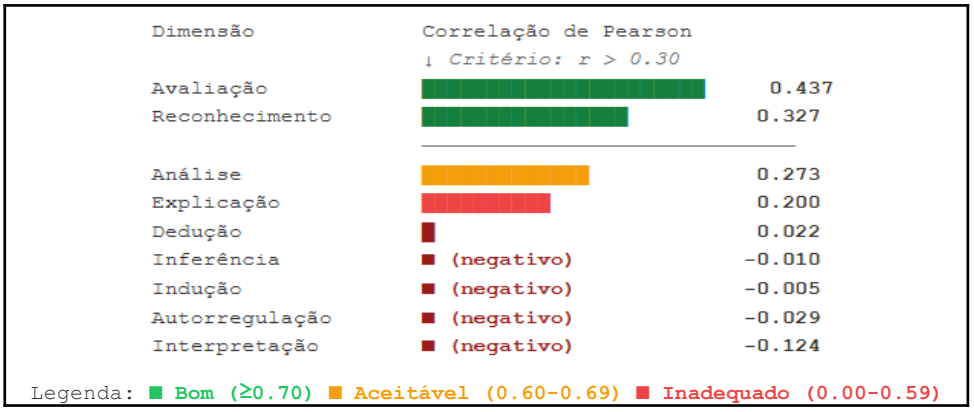

Figura 3. Validade convergente por dimensão

4.3. Quais são as evidências da validade de construto?

A validade de construto foi examinada através da correlação item-total corrigida (DeVellis, 2017), utilizando o critério de r.drop > 0.30 como adequado (DeVellis, 2017). Das 27 questões fechadas analisadas, 15 itens (55.6%) apresentaram correlações adequadas, variando de 0.311 a 0.757 (Figura 4).

Os itens com melhor desempenho foram relacionados às dimensões de "Explicação" e "Interpretação", com destaque para a questão "*Explico meu código de forma clara para qualquer pessoa*" (r = 0.757) e "*Identifico a função principal de um programa ao ler seu código*" (r = 0.647). Itens das dimensões "Autorregulação" também demonstraram correlações consistentes. Por outro lado, 12 itens apresentaram correlações abaixo do critério estabelecido, concentrando-se principalmente nas dimensões "Análise", "Inferência" e "Indução". Dois itens da dimensão "Análise" apresentaram correlações especialmente baixas (r = 0.090 e r = 0.177), corroborando os problemas identificados na análise de confiabilidade.

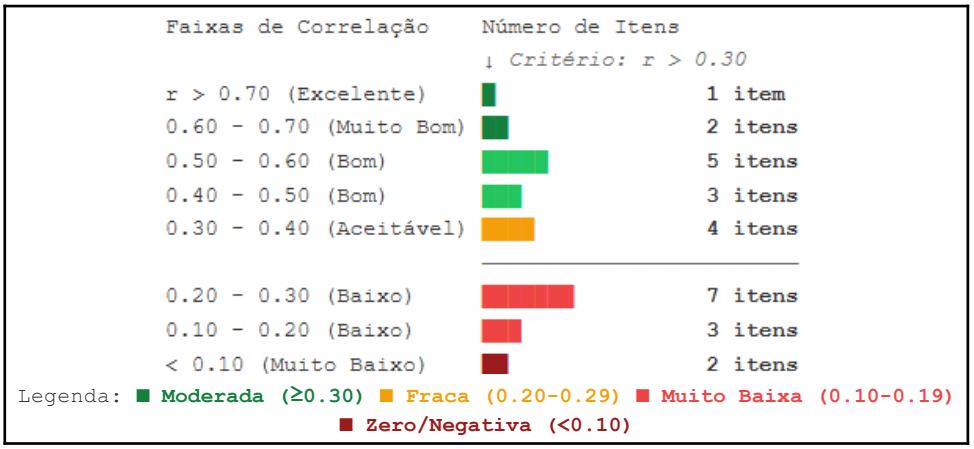

Figura 4. Correlações entre Questões Abertas (Rubrica) × Fechadas (Likert)

4.4. Quais são as evidências da estrutura fatorial?

A estrutura fatorial foi investigada através da Análise de Correspondência Múltipla (ACM) (Greenacre, 2017) aplicada às respostas das questões abertas, categorizadas conforme a rubrica analítica (escala de 1-Inadequado a 5-Excelente). A ACM das respostas abertas indicou uma estrutura fatorial fragmentada, com apenas 21,2% da

variância explicada pelas duas primeiras dimensões. Isso mostra que as nove dimensões teóricas do pensamento crítico não se agruparam empiricamente de forma clara, sugerindo alta dispersão e ausência de padrões consistentes nos dados. Observou-se ainda polarização nas respostas, com concentração nos extremos da rubrica e poucos casos nos níveis intermediários. Esses resultados podem estar relacionados à dificuldade dos critérios intermediários em discriminar graduações de competência, à heterogeneidade da amostra ou à sobreposição conceitual entre algumas dimensões no contexto da programação

4.5. Análise Qualitativa das Respostas Abertas

As respostas abertas foram avaliadas com base em uma rubrica analítica construída conforme o *framework Evidence-Centered Design* (*ECD*), considerando cinco níveis de desempenho: Inadequado, Razoável, Satisfatório, Bom e Excelente. A análise qualitativa teve como foco identificar evidências explícitas das habilidades de pensamento crítico previstas no modelo teórico (Facione, 1990; Yeh, 2003), tais como análise, inferência, explicação, interpretação, entre outras.

Os dados dessa análise revelaram uma polarização no desempenho dos estudantes, com maior concentração nos extremos da rubrica (“Inadequado” e “Excelente”), e poucos casos nos níveis intermediários, corroborando com os achados da Análise de Correspondência Múltipla. Esse padrão sugere que a rubrica precisa de ajustes nos critérios intermediários para aumentar a capacidade discriminativa. Além disso, foi possível observar que as habilidades mais evidentes nas respostas abertas foram “Interpretação” e “Explicação”, expressas por meio de justificativas claras nas respostas dos estudantes. Por outro lado, dimensões como “Análise” e “Inferência” apresentaram baixa frequência ou formulações genéricas, dificultando a identificação de indicadores consistentes nessas categorias.

4.6 Ameaças a validade

Foram identificadas duas principais ameaças à validade do instrumento. A primeira é a discrepância entre médias elevadas nas questões fechadas e baixas nas abertas, sugerindo possível viés de desejabilidade social ou desalinhamento entre itens e critérios de avaliação. A segunda refere-se à inconsistência entre dimensões: enquanto “Interpretação” e “Autorregulação” apresentaram bons índices psicométricos, outras requerem revisão. Adicionalmente, o tamanho amostral reduzido (N=29) e a aplicação em um único contexto limitam a generalização dos resultados.

5. Discussão e Conclusões

Este estudo piloto avaliou a adequação do ProgCTQ para mensurar o pensamento crítico no ensino de programação. Os resultados indicam que o instrumento apresenta confiabilidade geral satisfatória ($\alpha = 0{,}813$) e bom desempenho em dimensões como “Interpretação” e “Autorregulação”, sugerindo que a fundamentação teórica utilizada é pertinente para o contexto da computação.

Por outro lado, limitações importantes foram identificadas. A validade convergente foi baixa ($r = 0{,}121$) e a estrutura fatorial mostrou-se fragmentada, com apenas 21,2% da variância explicada. Isso indica que o instrumento, em sua forma atual,

pode estar capturando diferentes aspectos do pensamento crítico por meio das modalidades de avaliação abertas e fechadas. A discrepância entre os formatos sugere que as questões fechadas tendem a refletir autopercepções, enquanto as abertas evidenciam a capacidade de argumentação prática dos estudantes. Além disso, a polarização das respostas nos extremos da rubrica aponta para a necessidade de revisar e calibrar os critérios intermediários, de modo a melhorar a discriminação entre diferentes níveis de competência.

Esses achados reforçam a importância de ajustes metodológicos antes de uma aplicação em larga escala.

Apesar das limitações, o ProgCTQ representa um avanço relevante para a avaliação do pensamento crítico em computação, oferecendo uma base metodológica para o desenvolvimento de instrumentos com qualidade, confiáveis e contextualizados.

Direções futuras. Como trabalhos futuros, serão realizados: (i) reformulação dos itens das dimensões com baixa confiabilidade, baseada em análise qualitativa das respostas coletadas; (ii) revisão dos critérios da rubrica para aumentar a discriminação entre níveis de desempenho intermediários; (iii) aplicação em amostra ampliada (N≥100) para análises psicométricas com maior poder estatístico; (iv) validação cruzada em diferentes contextos institucionais; e (v) investigação de modelos fatoriais alternativos que possam melhor representar a estrutura empírica dos dados.

5.1. Síntese das Evidências Psicométricas do ProgCTQ

Uma síntese das evidências pode ser observada na Tabela 1.

Tabela 1. Síntese das Evidências Psicométricas do ProgCTQ

Evidência	Técnica	Nível	Resultado	Interpretação ■ Aprovado ▲ A observar
Confiabilidade	Alfa de Cronbach	Geral	$\alpha = 0.813$	■ Boa
		Por dimensão	$\alpha = -0.636$ a 0.727	▲ Variável
Validade Convergente	Correlação Pearson	Escalas x Rubrica	$\bar{r} = 0.121$	▲ Baixa
Validade de Construto	Correlação item-total	Escalas	55.6% itens ≥ 0.30	▲ Mista
Estrutura Fatorial	ACM	Rubrica	21.2% variância	▲ Fragmentada

Agradecimento

Os autores agradecem aos estudantes que participaram voluntariamente da pesquisa e à instituição de ensino que autorizou a coleta de dados.

Referências

Arndt, D. M., Martins, R. M., & Hauck, J. C. R. (2025a). 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).

Arndt, D. M., Martins, R. M., & Hauck, J. C. R. (2025b). Avaliação do desenvolvimento do pensamento crítico no ensino de programação para estudantes do ensino técnico e superior: Uma proposta de modelo. Manuscrito submetido para apresentação no Workshop sobre Educação em Computação (WEI). https://doi.org/10.5753/wei.2025.7944

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.

DeVellis, R. F. (2017). Scale development: Theory and applications (4th ed.). SAGE Publications.

Durak, H. (2020). The effects of using different tools in programming teaching of secondary school students on engagement, computational thinking and reflective thinking skills for problem solving. Technology, Knowledge and Learning, 25, 179–195.

Facione, P. A. (1990) California critical thinking skills test: CCTST. California Academic Press.

Greenacre, M. (2017). Correspondence analysis in practice (3rd ed.). Chapman and Hall/CRC.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.

In, J. (2017). Introduction of a pilot study. Nursing Research, 22(5), e1416.

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.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

OCDE. PISA 2018 Assessment and Analytical Framework: Reading, Mathematics and Science. Paris: OECD Publishing, 2019.

Pontual Falcão, T., & França, R. S. de. (2021). Computational thinking goes to school: Implications for teacher education in Brazil. Revista Brasileira de Informática na Educação, 29, 1158–1177.

Saunders, M., Lewis, P., & Thornhill, A. (2019). Research methods for business students (8th ed.). Pearson Education Limited.

Seeratan, K. L. & Mislevy, R. J. (2008). Design patterns for assessing internal knowledge representations. SRI International, Menlo Park. (PADI Technical Report 22).

Silva, D. E. S; Sobrinho, M. C.; Valentim, N. M. C. (2023) Development of 21st-century skills and competencies in high school students through the interactive e-books creation. RBIE - Revista Brasileira de Informática na Educação, v. 31, p. 971–1004.

UNESCO. Education for Sustainable Development Goals: Learning Objectives. Paris: UNESCO Publishing, 2017.

Yeh, Y.-C. (2003). Critical Thinking Test–Level I (CTT-I). Psychological Publishing