

Research Article

Cite this article: Gómez-Velasco, N. Y., Jiménez-Gonzalez, A. E., & Ortiz-Padilla, M. E. (2026). Alternative Approaches to Validate a Critical Thinking Measurement Instrument for University Students. *Educational Process: International Journal*, 22, e2026047. <https://doi.org/10.22521/edupij.2026.22.47>

Received September 1, 2025

Accepted October 29, 2025

Keywords: Critical thinking, validity metrics, robust alternative, HCTAES-Colombia

Author for correspondence:

Myriam Esther Ortiz-Padilla

 myriam.ortiz@unisimon.edu.co

 Universidad Simón Bolívar, Colombia

Alternative Approaches to Validate a Critical Thinking Measurement Instrument for University Students

Nubia-Yaneth Gómez-Velasco , Ana Emilce Jiménez-González , Myriam Esther Ortiz-Padilla 

Abstract

Background/purpose. The aim of this study is to present classical and robust alternatives for the validation of an instrument that assesses critical thinking (CP). The Chilean adaptation of the Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES) test was analyzed using three factorial models and a short-version model, with two estimation methods: maximum likelihood (ML) and weighted least squares (WLSMV). Internal consistency was examined with point and interval estimation for different reliability coefficients.

Materials/methods. The sample consisted of 214 students from a Colombian university. The instrument used was the Chilean adaptation of the Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES), which assesses five dimensions of critical thinking through open-ended and closed-ended items. For the psychometric analysis, three-factor models and an abbreviated version were considered, using Maximum Likelihood Estimation (MLE) and Weighted Least Squares (WLS) as estimation methods. Internal consistency was evaluated using reliability coefficients with confidence intervals, while construct validity was examined using Confirmatory Factor Analysis (CFA), complemented by convergent and discriminant validity indicators.

Results. The results indicate that the complete test and the open-ended question factor exhibit satisfactory internal consistency ($\omega=0.78$ and $\omega=0.79$, respectively). The confirmatory factor analysis using the robust WLSMV estimator yielded better fit across the absolute, incremental, and parsimony fit indices, corroborating Halpern's theoretical model.

Conclusion. It is concluded that the configural and open-format models retain adequate psychometric properties, i.e., good fit and stable factor configurations across the evaluated structures. The study contributes methodological elements to the line of validation and evidence of psychometric properties for PC, or serves as a reference for other analogous constructs.



OPEN ACCESS

© The Author(s), 2026. This is an Open Access article, distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction, provided the original article is properly cited.

1. Introduction

Critical thinking (CT) is a cross-cutting skill in higher education, as it is linked to comprehension, logical reasoning, problem-solving, reflective judgment, and informed decision-making in the complex contexts of modern life (Bezanilla, 2016). These skills are recognized as essential for training professionals who can analyze social, economic, and environmental issues, thereby contributing significantly to the construction of a more reflective and socially responsible society (Indrašienė et al., 2021). Their development and assessment are important objectives in university educational processes and pose a challenge for educators seeking to promote higher-order learning (Taseer et al., 2023).

The Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES) test has established itself as a relevant tool for assessing CT, combining open-ended and closed-ended questions that allow for inquiry not only into correct answers but also into underlying cognitive processes (Halpern, 2016). Unlike other instruments that focus exclusively on multiple-choice items, the HCTAES assesses the application of skills such as hypothesis testing, verbal reasoning, argument analysis, decision-making, and uncertainty management, which are fundamental aspects for measuring the complexity of thinking (Halpern, 2014).

The present study addresses the problem of insufficient psychometric evidence for validating the HCTAES in the Colombian population, which limits its reliable use in educational research and university teaching practice. Within this framework, the main objective was to present classic and robust alternatives for validating the instrument, exploring three factorial models and a short version differentiated between open and closed questions, using both classic and robust estimation methods.

Although the Halpern Critical Thinking Assessment (HCTAES) has demonstrated validity and reliability across various international contexts (Halpern, 2016; Butler, 2024), in Colombia there are still no systematic studies analyzing its factorial structure and psychometric properties using robust estimation methods. Most previous research on critical thinking in the Colombian university population has relied on self-report instruments or partial adaptations of foreign scales without full validation, which limits comparability of results and the application of theoretical models consistent with empirical evidence. In this sense, the present study addresses an empirical and methodological gap by providing a structural validity analysis of the HCTAES using descriptive and inferential statistical procedures, thereby contributing to the consolidation of a relevant measure applicable across different contexts for evaluating critical thinking in higher education.

The results indicate that the most appropriate estimation method for the Colombian case is WLSMV, as it offers greater robustness to non-normality and is suitable for handling categorical variables. Likewise, it was evident that model M1, which integrates open-ended and closed-ended questions, presents the best levels of reliability and validity. As for the short version, using the open format and adjusting items with lower factor loadings is recommended. These findings make two important contributions: on the one hand, they broaden the methodological discussion on the validation of instruments by integrating both classical and robust methods in the same study; and on the other hand, they strengthen the line of research on critical thinking in Colombia by offering solid psychometric evidence of the HCTAES in its five dimensions.

Finally, this work invites reflection on the instrument's pedagogical potential beyond its evaluative nature, as its open-ended questions allow recognition of students' reasoning and create opportunities for argumentation and the confrontation of ideas in the classroom. In this sense, the validation of the test not only constitutes an advance in technical terms but also a contribution to the design of educational experiences aimed at developing critical thinking as a core component of university education.

2. Literature Review

Critical thinking (CT), recognized as one of the fundamental competencies in higher education, as it is linked to problem solving, informed decision making, and analytical skills in social and professional contexts, continues to motivate research interest in different parts of the world (Tarasova et al., 2025; Campo et al., 2023; Rivas et al., 2023).

In relation to international evidence on the validity of the Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES), studies conducted in Europe report evidence of adequate validity and reliability of this test in different contexts and populations (De Bie et al., 2015; Butler et al., 2012; Rodrigues et al., 2018; Nieto et al., 2009). These studies have shown that the HCTAES can discriminate among levels of critical thinking performance and that its psychometric properties remain stable across various educational settings. Additional research in the United States and Asia has confirmed the usefulness of the HCTAES as a valid measure of CT, demonstrating acceptable internal consistency and replicable factor structures across cultures (Ku, 2009; Liu et al., 2014; Tiruneh et al., 2016). However, most of these studies do not report exploring other factorial models that integrate open-ended and closed-ended questions and examine the different correlations between them, nor do they explore short versions of the instrument.

In the Latin American context, and particularly in Colombia, evidence on the validity of the HCTAES is even more limited and fragmentary. Some recent efforts have addressed this gap, highlighting the need for more comprehensive validation processes to ensure the instrument's cultural and psychometric relevance (Rodriguez et al., 2025; Maturana-Moreno & Lombo-Sánchez, 2020; Ortega-Quevedo et al., 2020; Saldaña et al., 2022). These initial studies have shown significant progress in linguistic adaptation. However, they also warn of the need to explore alternative factor models that account for the particularities of Colombian educational contexts (Gómez-Velasco et al., 2024).

In addition, studies published in high-impact journals indicate that the construct validity of critical thinking tests depends not only on linguistic adaptation but also on cultural and pedagogical coherence in each country (Ennis, 2018; Niu et al., 2013). Likewise, other studies have shown that tests that include both open-ended and closed-ended questions, such as the HCTAES, have advantages over multiple-choice instruments alone in capturing a broader range of cognitive skills (Liu et al., 2020; Marin & Halpern, 2011). In this context, the application of relevant statistical methods, such as confirmatory factor analysis and robust estimators, especially for ordinal variables, is still required to obtain more accurate results in psychometric validation (Hair et al., 2021; Brown, 2015).

3. Methodology

3.1. Type of study

The study was carried out with a non-experimental descriptive quantitative methodological approach, with an instrumental design, taking into account that the psychometric properties of a test to measure critical thinking are analyzed (Ato et al., 2013), with adaptations in linguistic and cultural aspects following the guidelines of Muñiz et al. (2013).

3.2. Participants

A sample of 214 students from the Psychology program at a private university in Barranquilla, Colombia, was compiled, representing the second (33%), fifth (46%), and eighth (21%) semesters, with a higher percentage of females (84%). The average age was 22 years ($SD = 2.3$), and the range was 18-28 years. According to Ferrando and Anguiano (2010), an adequate sample size for obtaining accurate estimates in the adjusted model should exceed 200.

The selection of participants was made through non-probabilistic purposive sampling, aimed at ensuring the inclusion of students representing different stages of the formative process in Psychology. The second, fifth, and eighth semesters were considered since they correspond to different stages of academic and professional development. This strategy made it possible to capture diversity in perceptions and formative experiences, ensuring a broad coverage of the program.

It should be noted that the sample does not pretend to be representative of the general population of university students and an imbalance is recognized in the distribution by sex, with 84% of women; however, this work is conceived as an exploratory study focused on evaluating the relevance, internal consistency and psychometric behavior of the instrument in a particular context, which could serve as a basis for future applications in more diverse and representative samples.

3.3. Instrument

The instrument submitted for validation was the HCTAES (Halpern Critical Thinking Assessment Using Everyday Situations), designed by Halpern (1998, 2016), adapted to Spanish by Nieto et al. (2009), and adapted to the Chilean population by Fuentes (2010). This test evaluates five critical thinking skills: Verbal Reasoning (VR), Argument Analysis (AA), Hypothesis Testing (HT), Probability and Uncertainty (PU), and Decision Making and Problem Solving (DMPS). It consists of 25 everyday situations with open-ended and closed-ended questions. For each situation, the individual first provides an open-ended response and then selects the option with which he or she most identifies in the closed-ended format (Halpern, 1998, 2016).

The evaluation of the responses was conducted according to the criteria defined in the HCTA manual, which specifies maximum scores for the total, for each ability, and for each question format. The psychometric properties reported by the author indicate acceptable reliability: open format ($\alpha=0.78$), closed format ($\alpha=0.68$), and overall ($\alpha=0.82$), with a factorial structure of two correlated latent variables and congruence between skills (Halpern, 2016).

3.4. Procedure

The author of the Chilean version of the HCTAES test used in this study was contacted and approved, given that its distribution was not free. Initially, a pilot test was conducted with 17 students and was evaluated by expert judges. Linguistic adaptation was proposed in the context of the students, using terms that could pose comprehension difficulties, among them: "posture" for points of view, "consensus" for agreement between the parties, and "tertulia" for conversation. The administration of the test was conducted in a pencil-and-paper format, with voluntary participation and informed consent, in accordance with the rules on habeas data and confidentiality. Instructions on the instrument and the research's academic purposes were explained to them. The development of the questionnaire took between 90 and 120 minutes.

The research protocol was entered and approved in the institutional research management system of the Universidad Pedagógica y Tecnológica de Colombia (SGI-UPTC) under project code SGI 2917, which certifies compliance with the ethical processes required by the institution.

3.5. Statistical analysis

The statistical analysis was conducted using descriptive measures and psychometric analyses. The descriptive exploration was carried out by summing the items for each ability, using an open and closed-question format, and calculating the Pearson correlations between abilities. Mardia's multivariate normality test suggests that the distribution is mesokurtic (skewness, $p = .001$; kurtosis, $p = .97$); however, the normality assumption is not met (Cain et al., 2017), which implies the use of robust estimation methods.

Internal consistency was assessed using point and interval estimates (95%) for different reliability coefficients; Cronbach's alpha was calculated because it is a measure reported in the literature that allows comparisons. The values of the lower limit of the confidence interval were greater than or equal to 0.70 and are therefore considered evidence of acceptable reliability. Due to the limitations of the alpha coefficient regarding the uniqueness and multiplicity of item responses, increasing the probability of not meeting the tau equivalence assumption (Dunn et al., 2014), the coefficient Omega of McDonald's (ω) (Sijtsma, 2009) was calculated with an acceptable result for values between 0.7 and 0.9 (Ventura & Caycho, 2017). Additionally, the Greatest Lower Bound (GLB) coefficient was calculated, an appropriate estimator for a high proportion of asymmetric items. In this case, a GLB value of 0.8 sets the lower limit of the interval, implying that the true reliability lies between 0.8 and 1 (Peters, 2018).

Total test-retest reliability was determined for each ability and for each question format. On the grounds that the item scores are not equal across the HCTAES skills, the coefficients were calculated using both the original and standardized scores.

The validation of this test was performed using AFC, employing the SEM (Structural Equation Modeling) technique, which provides adjustment alternatives and is recommended when items do not represent perfect indicators and have low, non-zero weights in other factors (Marsh et al., 2020). The SEM technique is suggested for testing theories and validating measurement instruments (Taseer et al., 2023).

For comparative purposes, the classical maximum likelihood estimation method MV and the robust weighted least squares method WLSMV were used, with WLSMV being the recommended method in the case of non-compliance with the model assumptions (Lai, 2018), which, for the purpose of validating an instrument with the characteristics of the HCTAES, made possible a favorable spectrum for a good fit.

The goodness of fit of the model was evaluated with global fit measures, among which are: Chi-square statistic (χ^2 , $p < .05$) (Wood, 2008), Root Mean Squares Error Approximation index (RMSEA < 0.05) (Mulaik, 2009); in the case of incremental fit measures were: normed fit index- NFI; comparative fit index-CFI- Comparative Fit Index and the Tucker-Lewis index-TLI (Xia & Yang, 2019), measures with values that are usually bounded between 0 and 1, where 1 indicates that the fitted model is faithful to the theoretical model.

As for the parsimony fit measures, the PNFI Normalized Fit Index was used, which relates the constructs to the theory that underlies them, the closer it is to 1 the greater their relationship; PGFI Parsimony Goodness of Fit Index, which constitutes a modification of the GFI and considers the degrees of freedom available to test the model, where the magnitudes considered acceptable are in the range of .5 to .7 (Mulaik, 2009).

Although Halpern's (1998) proposal consists of 5 skills each with 10 items, 5 with a closed response format and 5 with an open or constructed response format, for the validation of the formulated model (M1, M2 and M3) two factors correspond to the response formats (open or closed) and the items that comprise them correspond to the five skills.

The analysis of the M1 model is presented in accordance with the theoretical proposal of Halpern (1998), which hypothesizes that the scale scores on the five skills with constructed response question load on the latent factor "PC Open-ended Question" (OQo) and the scores on the five skills with multiple choice question load on the latent factor "PC Closed-ended Question" (OQc); with correlation between factors and between open-ended skills with closed-ended skills, as reported in Halpern's (2016) manual and Rodrigues et al, (2018).

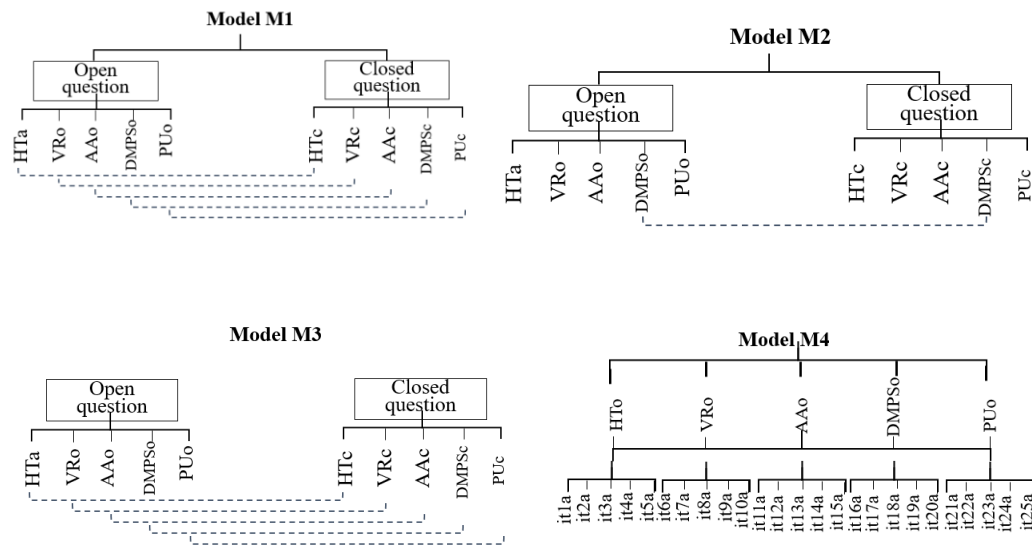


Figure 1. Validated critical thinking models

Model M2 differs from M1 only in the inclusion of the correlation between the skill: problem solving (TDRP), in open and closed formats, given the significant correlation between them. Model M3 differs from M1 in that the latent factors are treated as completely separate traits, with the correlation between them set to zero. For each model, estimates were obtained using the classical (MV) and robust (WLSMV) methods, denoted as M1(MV), M2(MV), M3(MV), M1(WLSMV), M2(WLSMV), and M3(WLSMV).

With the factor loadings obtained, the composite reliability coefficient was calculated; values greater than .7 are considered acceptable (Green & Yang, 2015). To enable comparisons, the calculation was also performed using the data from the reference studies. The convergent validity of each factor was studied with the mean variance extracted SMV, which is acceptable if $SMV > .5$ (Jöreskog et al., 2016). The discriminant validity between factors was assessed using two criteria: Fornell-Larcker, which is satisfied if VME exceeds the square of the correlation between factors. The heterotrait-monotrait ratio (HTMT) of correlations is considered met if $HTMT < .85$ (Henseler et al., 2015).

Additionally, the psychometric properties of a model that provides a short version of the test, in an open-ended question format (M4: five factors and a total of 25 items), were explored. A summary of the techniques and measures used in the instrument's analysis is presented in Figure 2.

The analyses presented in the results were performed using the libraries Lavan, psych, semTools, MBESS, and reshape2 in the statistical program R version 4.0.2, and the program JASP version 0.17.2. The figure of the standardized factor loadings (Figure 2) was elaborated with the program SPSS-AMOS 24.

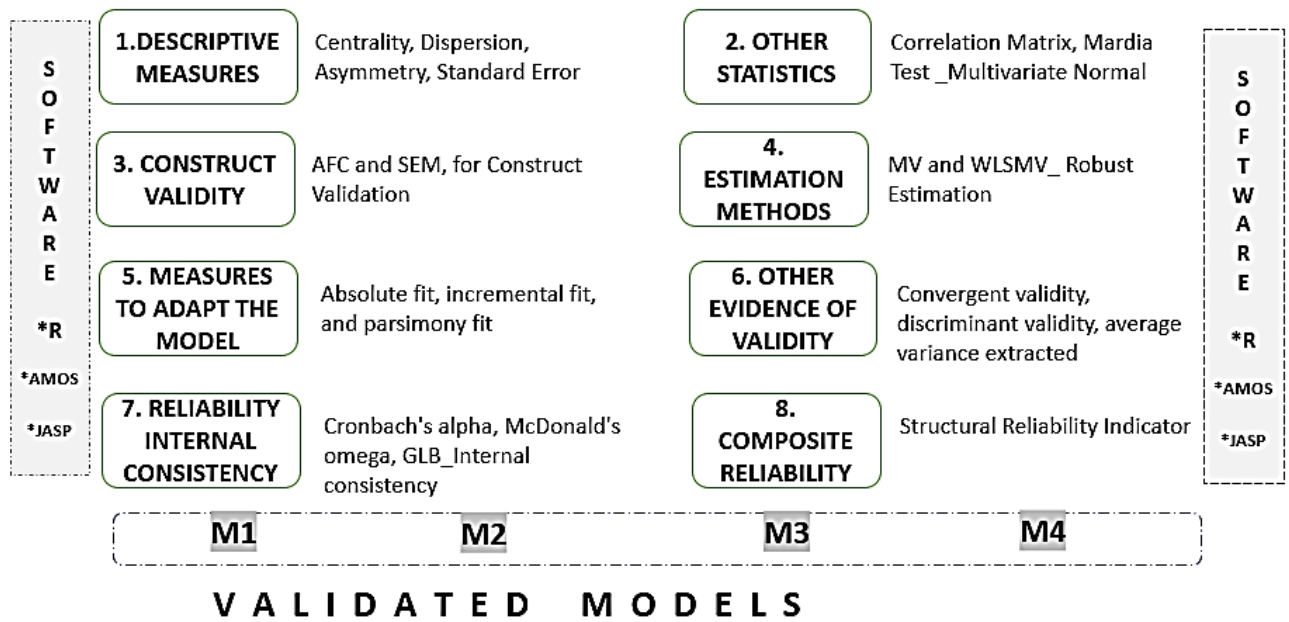


Figure 2. Techniques and measures for validation and reliability. Critical Thinking Test-HCTAES

4. Results

4.1. Descriptive analysis

The descriptive measures for each ability and for the total score are presented in Table 1, considering the open (a) and closed (c) formats, also specifying the maximum possible score in each ability, for the test in general (50 items) and in each format, according to the test manual.

Table 1. Descriptive Measures by Abilities and Type of Questions

PME	Skills (#items)	M(DE)	Min;Máx	As	Cu	EE
46	HT(10)	19.1 (5.1)	6; 30	.03	-.51	.35
19	HTo(5)	7.1 (3.5)	0; 17	.13	-.26	.24
27	HTc(5)	12.1 (3.1)	3; 19	-.45	.01	.21
22	VR(10)	8.5 (2.8)	2; 17	.25	-.32	.19
15	VRo(5)	5.2 (2.2)	1; 12	.54	-.04	.15
7	VRc(5)	3.3 (1.3)	0; 7	-.18	-.39	.09
41	AA(10)	16.5 (4.9)	5; 28	.05	-.58	.33
22	AAo(5)	6.9 (3.7)	0; 18	.39	-.39	.25
19	AAc(5)	9.6 (2.9)	2; 16	-.16	-.38	.20
24	PU(10)	9.8 (4.4)	0; 20	.20	-.65	.30
17	PUo(5)	6.9 (3.9)	0; 16	.24	-.73	.27
7	Plc(5)	2.9 (1.3)	0; 6	.05	-.46	.09
61	DMPS(10)	33.3 (8.9)	0; 49	-1.38	2.98	.61
22	DMPSo(5)	10.1 (4.7)	0; 19	-.28	-.62	.32
39	DMPSc(5)	23.2 (5.8)	0; 33	-1.83	4.49	.40
194	TOTAL(50)	87.3 (19.7)	39; 132	-.40	-.30	1.35
95	Ta(25)	36.2 (13.3)	6; 66	-.08	-.83	.91
99	Tc(25)	51.1 (9.9)	18; 72	-.95	.79	.68

Note: PME: Maximum Expected Score. M: Mean. SD: Standard deviation. As: Asymmetry. Cu: Kurtosis. SE: Standard error.

It can be seen that for each skill, the scores of the open part are lower with respect to its PME than the scores of the closed part with respect to its respective PME. The analysis of skewness and kurtosis across the different skills falls within an acceptable range (-1.5 to 1.5) (Ferrando & Anguiano, 2010), except for TDRP.

A correlational study was conducted on the different skills in open and closed formats (Table 2). The skills in the open format are correlated, with correlations ranging from .37 to .57; the skills PIa and TDRPa show the highest correlation, followed by PHa and Pla. High correlations are observed between each open-format skill and the total (corrected) score for this format, indicating that the open-format skills contribute to measuring PC.

Table 2. Table of correlations between skills

Ability	HTo	VRo	AAo	PUo	DMPSo	Ta	HTc	VRc	AAc	PUc	DMPSc
HTo											
VRo	.41**										
AAo	.40**	.37**									
PUo	.49**	.39**	.47**								
DMPSo	.42**	.41**	.43**	.57**							
Ta	.54**	.50**	.54**	.66**	.61**						
HTc	.23*	.19*	.12	.19*	.14*	.23**					
VRc	.28*	.23*	.17*	.19*	.27*	.29**	.33**				
AAc	.19*	.17*	.13	.25**	.26**	.26**	.28**	.17*			
PUc	.26**	.16*	.15*	.29**	.23**	.29**	.19*	.26**	.23**		
DMPSc	.14*	.11	.13	.26**	.36**	.34**	.34**	.29**	.27**	.19*	
Tc	.28*	.24**	.19*	.36**	.44**	.42**	.41**	.40**	.34**	.29**	.42**

Note: Correlation significant at .01**. Correlation significant at .05*.

With respect to the correlation between the skills in the closed format, we observe that they are low with values ranging between .19 and .34, the highest being between PHc and TDRPc; in the closed format the correlations between each skill and the corrected total are lower than those obtained in the open format, with the Probability and Uncertainty skill being the lowest (.29), that is, the one that contributes the least to measuring PC. In addition, the correlations between the open- and closed-format skills range from .13 to .29, indicating that all the skills contribute to measuring the construct and are not redundant. Finally, the correlation between the total open- and closed-format scores is also moderate (.423). It can be observed that the low correlations among the different skills support the test's multidimensionality and the independence of the factors.

4.2. Reliability Analysis

Table 3 presents the reliability analysis for each ability, each question format, and for the total. The different reliability indices (Cronbach's Alpha, Omega, and GLB) report values that are close to each other, with a tendency toward higher values for the GLB index.

Table 3. Reliability measures for each skill and the total

Ability	α Cronbach (IC95%)	Omega (IC95%)	GLB
HT	.51(.41, .62)a	.51(.41, .61)a	.67a
	.50(.39, .6)b	.49(.4, .6)b	.67b
VR	.35(.22, .48)a	.34(.19, .48)a	.49a
	.34(.21, .48)b	.35(.21, .49)b	.49b
AA	.42(.31, .54)a	.45(.34, .56)a	.6a
	.43(.32, .55)b	.39(.26, .53)b	.6b
PU	.55(.46, .65)a	.6(.52, .69)a	.67a
	.52(.42, .61)b	.53(.42, .61)b	.67b
DMPS	.78(.73, .82)a	.79(.75, .83)a	.85a
	.78(.74, .82)b	.78(.73, .82)b	.85b
Ta	.82(.78, .85)a	.82(.79, .86)a	.9a
	.81(.77, .85)b	.81(.77, .85)b	.9b
Tc	.73(.67, .79)a	.77(.72, .81)a	.82a
	.69(.63, .75)b	.69(.63, .75)b	.82b
Total	.84(.81, .87)a	.85(.82, .88)a	.83a
	.83(.8, .86)b	.83(.8, .86)b	.85b

Note: *a*unstandardized item. *b*standardized item.

The reliability of the different skills ranges from .34 to .79, as measured by both Cronbach's α and McDonald's ω . The lowest value was obtained in the RV skill, and the highest in TDRP. The open-ended questions reported satisfactory reliability value (α and $\omega > .8$) and the closed-ended questions acceptable reliability ($.7 < \alpha$ and $\omega < .8$). The test-retest reliability value is good, with a test total close to .85 (with the different coefficients). Likewise, it is observed that, both in the standardized and unstandardized data, the reliability values do not differ significantly.

4.3. Construct Validation

The content validation process was carried out using models M1, M2, and M3, which involve two factors, each with 5 items, corresponding to the skills: Factor 1: OQ open-ended questions (CQo) and Factor 2: PC closed-ended questions (CQc). The factor structure was analyzed using CFA.

Table 4 presents the global, incremental and parsimony fit measures, for the comparison of the three models under the classical maximum likelihood estimation (MV) and weighted maximum likelihood robust estimation (WLSMV) methods, denoted by: M1(MV), M2(MV), M3(MV) and M1(WLSMV), M2(WLSMV) and M3(WLSMV).

Table 4. Measures to assess model adequacy

MODELS	Overall fit			Adjustment incremental				Adjustment Parsimony		
	χ^2/df	p	RMSEA	NFI	RFI	IFI	TLI	CFI	PNFI	PGFI
M1(MV)	32.41/29	.30	0.02	0.93	0.89	0.99	0.98	0.99	0.60	0.51
M2(MV)	37.24/33	.28	0.02	0.92	0.89	0.99	0.98	0.99	0.68	0.58
M3(MV)	68.3/30	<.001	0.08	0.86	0.79	0.92	0.87	0.91	0.57	
M1(WLSMV)	15.26/29	.98	0.00	0.98	0.97	0.99	0.99	0.99	0.63	0.52
M2(WLSMV)	18.45/33	.98	0.00	0.98	0.97	0.99	0.99	0.99	0.73	0.59
M3(WLSMV)	167.7/30	<.001	0.15	0.79	0.80	0.82	0.72	0.82	0.53	

In general, the global fit, incremental fit, and parsimony measures yield better results for the M1 and M2 models with the WLSMV estimation method than with the MV method. Although, in terms of parsimony, the values are not very close to unity, they do exceed the value of .5, which implies an acceptable fit. For the M3 model the classical estimation method reported slightly better fit values.

When comparing the M1 and M2 models using the WLSMV method, there are no major differences in measures of global fit or incremental fit, but there are major changes in measures of parsimony, with favorable results for M2.

The measures indicate that the M1 model fits the data well. It should be noted that although M2 presents a similar fit, its structure links the correlation between the open and closed formats to only one skill. The fit measures show that model M3 does not meet the specifications for good fit. Figure 3 presents the standardized factor loadings and structural relationships in model M1, which corresponds to Halpern's (2016) configural model.

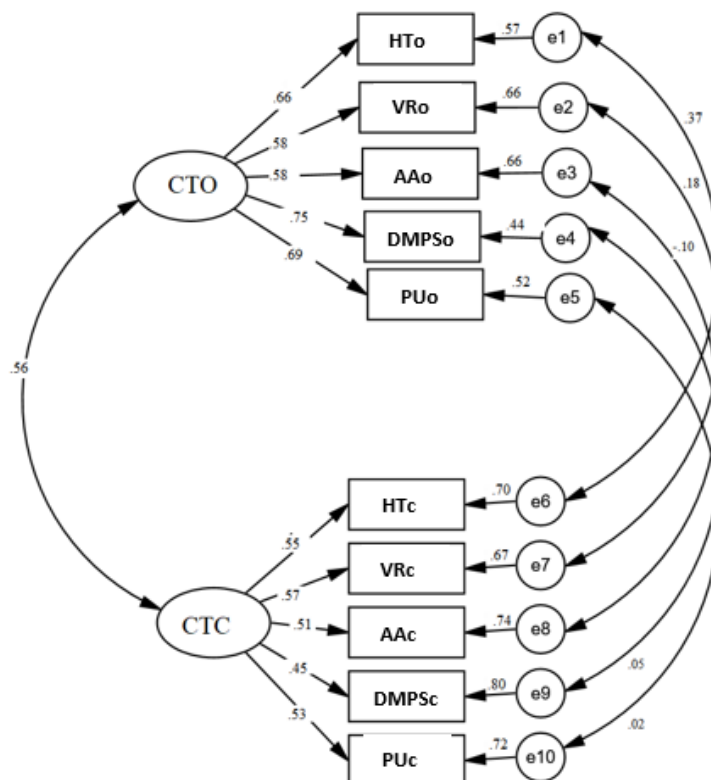


Figure 2. Standardized factor loadings of the HCTAES M1 Model, WLSMV method

The M1 model (Figure 3) shows that all skills have high factor loadings (greater than .4), i.e., the coefficients representing the degree of relationship between the construct (Critical Thinking) and their respective indicators (Skills assessed) are appropriate.

The internal consistency of M1, as measured by the composite reliability index, is adequate for the open-ended question (a) format (.79). At the same time, it is low for the closed-ended question (c) format (0.58). The factors did not exhibit convergent validity because values greater than .5 were not reached ($VMEa=.45$ and $VMEc=.28$), as evidenced by factor loadings not exceeding .7 for each factor. With the Formell-Larcker criterion, it is observed that the scale presents discriminant validity, with the values of the square root of the SMV being greater than the correlation between the factors ($.67 > .42$ and $.52 > .42$). With the criterion of the value of HTMT, the discriminant validity between the factors is confirmed ($HTMT=.59$) given that the value is less than .85.

4.4. Psychometric properties of the short version (M4)

The short version is a proposal for this study, which uses only the open format; hence, the M4 model consists of five factors with five items each. The results of the model exploration using the WLSMV estimation method are reported in Table 5. Fit that is improved by removing variables with saturations less than .3 (Jöreskog et al., 2016), denoted as M4(*).

Table 5. Measures for assessing M4 and M4(*) fit

MODELS	Overall fit			Fit Incremental				Adjustment Parsimony		
	χ^2/df	p	RMSEA	NFI	RFI	IFI	TLI	CFI	PNFI	PGFI
M4	305.98/265	.04	.03	.84	.82	.97	.97	.97	.74	.77
M4(*)	138.12/129	.27	.02	.90	.88	.99	.99	.99	.76	.73

Note: (*) Model with items eliminated due to saturation less than .3.

The short version proposal from the open-format questions (model M4) presented, in general, good fit in the three sets of measures, which are improved with M4 (*) in all incremental fit measures and some parsimony measures (PNFI), but not with respect to those of global fit, as in the RMSEA. The modification was obtained by eliminating 7 items with saturations less than .3 from the skills Hypothesis Testing (1), Argumentative Analysis (1), and Verbal Reasoning (5). This short version reports satisfactory reliability ($\alpha = 0.82$ and $\omega = 0.82$).

It was proposed to analyze the psychometric properties of the short version based on the closed-format questions; however, the model presented convergence problems, which is why a satisfactory solution was not found, possibly explained by the very low correlations within the same factor (Table 1) or by slightly high correlations of items from one factor with other factors (Basto & Pereira, 2012).

5. Discussion

The study validated the HCTAES instrument to assess critical thinking in university students, through three factorial models according to Halpern's (2016) configuration and a short version model, with the application of two estimation methods: the method that is classically applied, that is, maximum likelihood-MV and the proposal from a more robust method, called Weighted Least Squares-WLSMV, as reported in recent studies Acosta & Torres (2025).

The descriptive analysis reported a mean score for PC of 87.3, lower than those given in studies with Portuguese university students ($n=333$, mean 113) (Rodrigues et al., 2018), Dutch population ($n=240$, mean 108.23) (De Bie et al., 2015), or Spanish-speaking adult population ($n=355$, mean 106) (Halpern, 2016). The low average in Colombia may be associated with the educational gap between Latin American and European countries (Hill et al., 2008).

From the descriptive results, it was identified that the highest score in the closed-response items privileges the ability to recognize the good use of PC skills over the active use of these, as supported by Halpern (2016). Furthermore, the hypothesis that open-format scores are correlated with closed-format scores is corroborated by a statistically significant result.

The internal consistency for the overall test and for each question format was evaluated through three coefficients: Cronbach's α , McDonald's ω , and GLB, with the direct and standardized scores. With Cronbach's α , satisfactory values were obtained, close to the reference values in the manual ($\alpha=.88$; $\alpha_{open}=.83$ and $\alpha_{closed}=.77$) (Halpern, 2016) and comparable to the version adapted to the Dutch population ($\alpha=.75$; $\alpha_{open}=.61$ and $\alpha_{closed}=.64$) (De Bie et al., 2015). Reliability, as measured by coefficient ω , yielded higher values than those reported for the Portuguese population. ($\omega = .75$; $\omega_{open}=.70$ and $\omega_{closed}=.58$) (Rodrigues et al., 2018). Regarding the GLB coefficient, comparisons cannot be made because it is not reported in the reference studies; however, given that

it is an appropriate coefficient for tests with a high proportion of asymmetric items, as in this case, higher reliability values were found in all subscales of the test, compared to Cronbach's alpha and McDonald's ω .

Regarding reliability by skills, similar results were observed to the Spanish population (Nieto et al., 2009) and the Dutch population (De Bie et al., 2015) with lower values in the Verbal Reasoning skill ($\alpha < .40$). In this study it is highlighted, in contrast to previous studies, that the Problem Solving skill presents satisfactory reliability values ($\alpha > .70$).

The composite reliability for the open-ended question format was adequate (>0.7), whereas it was not for the closed-ended format. This result contrasts with the values calculated from the HCTA manual's factor loadings (Halpern, 2016), in which both formats yield acceptable composite reliability. Regarding the Portuguese and Dutch case (De Bie et al., 2015; Rodrigues et al., 2018), for both formats, the values are below .7.

The results of this study are consistent with international evidence supporting the validity of the Halpern Critical Thinking Assessment (HCTA) across different contexts. Several reviews and meta-analyses have highlighted its predictive and criterion-related validity. For example, Butler (2024) reports that higher scores on the HCTA are associated with fewer negative events in everyday life, which reinforces its ecological validity and relevance for assessing critical thinking in real-world situations. Similarly, cross-cultural studies have documented significant associations between HCTA scores and academic performance, as well as consistent generalization patterns across countries (e.g., comparisons between China and the United States), providing robust evidence of convergent validity. In this framework, the results of the present work contribute to extending such evidence in the Colombian context by offering an updated psychometric analysis that strengthens the instrument's use in national higher education.

For the construct validity of the M1 and M2 models the maximum likelihood estimation method loses generality and relevance according to the measures of absolute, incremental and parsimony fits, results that are notably improved with the robust WLSMV method that solves the problem of slight violations in the assumption of normality and offers better estimates for modeling categorical variables, furthermore, it performs acceptably with relatively small sample size (close to 200) and for two- and four-factor scales, independent of factor loading size (Liang & Yang, 2013).

The M1 model for the Colombian case presented fit measures higher than those reported in other studies (De Bie et al., 2015; Halpern, 2016; and Rodrigues et al., 2018). The model retained the two-factor structure and the open-ended and closed-ended question formats, with correlations similar to those reported in the HCTAES test manual (Halpern, 2016). The correlation of skills showed a more defined structure in the open format and lower values in the closed format, which could indicate a weaker structure in the closed format.

Model M2 loses strength compared to M1 by considering only the correlation between one skill and ignoring that the skill assessed by open-ended question and closed-ended question share the same question statement as indicated in Halpern (2016). Model M3 by presenting weak measures of fit is not recommended for PC analysis. The M4 model arises from the motivation to provide evidence of psychometric properties of a short version of the test, based on the open question format of the HCTAES instrument in the measurement of critical thinking, seeking to solve the limitation of the length and application time of the test (Halpern, 2016). This model presented satisfactory reliability levels and with global, incremental and parsimony adjustment measures, adequate under the WLSMV estimation method, therefore, its use can be recommended with the foresight of revising some items that showed low factor loadings. On the other hand, the test with open-ended question allows more opportunities for argumentation which favors constructs such as critical thinking (Saiz et al., 2022).

6. Conclusion

The exploration of the short version model with a closed-ended question format yielded acceptable levels of reliability; however, it did not pass the validity indexes due to convergence problems. Therefore, it is suggested to continue exploring the Halpern test with this question format in other populations. It is concluded that for the adjustment models the recommended estimation method is WLSMV because it is more robust against the non-compliance of the normality assumption and appropriate for modeling categorical variables. Regarding the models that jointly include open and closed format, better results were obtained with M1 because in general it has better levels of reliability and validity for the case of Colombia. In relation to the proposal to have a short version of the HCTAES, the open format version is recommended, with the precaution of eliminating items with low factor loadings, as identified in this study.

From the results found in this research, two contributions to the academic research community are emphasized: initially, in the line of instrument validation, where a diversity of alternative and robust statistical measures are included in a single document, which can serve as a reference for other studies; proposing validation from classical methods and advancing to non-classical methods. On the other hand, it contributes to the evaluation of PC in Colombia, where this research validates the HCTAES instrument across its five dimensions. The test proved to be a useful diagnostic instrument, serving as a reference for reflection and analysis of intervention processes in the classroom, rather than limiting itself to its results, both for teachers and for programs with this purpose. The students' answers to the open questions and the examination of their logic serve as a tool for creating dialectical exercise scenarios that confront arguments to achieve more coherent reasoning. As limitations of this study, the need for an appropriate space and time is observed, which favor the student's concentration to obtain results that account for their thinking, given that in this population a longer time than the average time proposed in the initial test was required, which could suggest great complexity in the problems posed, as well as inattention due to the length of the test.

7. Suggestion

Given the importance of positive provisions for developing critical thinking, future studies should expand the sample to include other population groups (students from other areas of knowledge and different regions of the country) to obtain more powerful estimates, validate linguistic adaptations, and apply the digitized test to facilitate evaluation. In addition, research into relationships with other constructs, such as academic performance or tests that measure disposition, could guide learning processes, addressing the need to integrate, into professional training, the motivation to learn to learn, to think critically about reality, and future social responsibility. In this sense, structuring classroom experiences that enable argumentation, confrontation, and dialectical exercise will add value to educational initiatives.

Declarations

Author Contributions. Authors Nubia Gómez-Velasco and Ana Emilce Jiménez: statistical analysis, manuscript writing. Author Mirian Ortiz: conception of the initial project idea, application of the instrument, and construction of the database.

All three authors participated in the general review of the document. All authors have read and approved the final version of the article.

Conflicts of Interest. The authors declare that there are no conflicts of interest in the conduct and publication of this study.

Funding. This project received financial support from the Pedagogical and Technological University of Colombia (UPTC).

Ethical Approval. The study respected ethical principles. The instrument was administered with participants' informed consent. The project was approved by the Ethics Committee of the Pedagogical and Technological University of Colombia.

Data Availability Statement. The data supporting the results of this study are available upon reasonable request to the corresponding author. We do not have authorization to share the instrument, as the creators of the Halpern instrument must endorse it.

Acknowledgments. The authors express their gratitude to the Pedagogical and Technological University of Colombia and the Simón Bolívar University of Barranquilla, Colombia, for their institutional support. We also thank the participants who made the data collection possible.

References

- Acosta Jiménez, J. C., & Torres Rojas, I. S. (2025). Validation of an instrument using modern psychometric techniques for the measurement of quantitative reasoning. *Caracas Medical Gazette*, 133.
- Anggraeni, D. M., Prahani, B. K., Suprpto, N., Shofiyah, N., & Jatmiko, B. (2023). Systematic review of problem based learning research in fostering critical thinking skills. *Thinking Skills and Creativity*, 49, 101334. <https://doi.org/10.1016/j.tsc.2023.101334>
- Ato, M., López-García, J. J., & Benavente, A. (2013). A classification system for research designs in psychology. *Anales de Psicología*, 29(3), 1038–1059. <https://doi.org/10.6018/analesps.29.3.178511>
- Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software*, 46(4), 1–29. <https://doi.org/10.18637/jss.v046.i04>
- Bezanilla, M. J., Domínguez, H. G., & Ruiz, M. P. (2021). Importance of teaching critical thinking in higher education and existing difficulties according to teachers' views. *REMIE: Multidisciplinary Journal of Educational Research*, 11(1), 20–48. <https://doi.org/10.4452/remie.2021.02>
- Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., & Almeida, L. S. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity*, 7(2), 112–121. <https://doi.org/10.1016/j.tsc.2012.04.001>
- Butler, H. A. (2024). Predicting everyday critical thinking: A review of critical thinking assessments. *Journal of Intelligence*, 12(2), 16.
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence, and estimation. *Behavior Research Methods*, 49(5), 1716–1735. <https://doi.org/10.3758/s13428-016-0814-1>
- Campo, L., Galindo-Domínguez, H., Bezanilla, M. J., Fernández-Nogueira, D., & Poblete, M. (2023). Methodologies for fostering critical thinking skills from university students' points of view. *Education Sciences*, 13(2), 132. <https://doi.org/10.3390/educsci13020132>
- De Bie, H., Wilhelm, P., & van der Meij, H. (2015). The Halpern Critical Thinking Assessment: Towards a Dutch appraisal of critical thinking. *Thinking Skills and Creativity*, 17, 33–44. <https://doi.org/10.1016/j.tsc.2015.04.001>
- Dunn, T. J., Baguley, T., & Brunsdon, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399–412. <https://doi.org/10.1111/bjop.12046>

- Ferrando, P. J., & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología [Factor analysis as a research technique in psychology]. *Papeles del Psicólogo*, 31(1), 18–33. <https://www.redalyc.org/articulo.oa?id=77812441003>
- Fuentes, C. (2010). *Halpern's Critical Thinking Test, version adapted to the Chilean context* [Manuscript no publicado]. Universidad Diego Portales.
- Furr, R. M. (2017). *Psychometrics: An introduction* (3rd ed.). SAGE Publications.
- Gómez-Velasco NY, Jiménez-González AE, Ortiz-Padilla ME. (2024). Methodological proposal for analyzing critical thinking test scores as a tool for development in higher education. *Journal of Infrastructure, Policy and Development*. 8(15): 10089. <https://doi.org/10.24294/jipd10089>
- Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, 34(4), 14–20. <https://doi.org/10.1111/emip.12100>
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449–455. <https://doi.org/10.1037/0003-066X.53.4.449>
- Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.
- Halpern, D. F. (2016). *Halpern Critical Thinking Assessment: Test manual HCTA*, Version 51. Schuhfried GmbH.
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177. <https://doi.org/10.1111/j.1750-8606.2008.00061.x>
- Indrašienė, V., Jegelevičienė, V., Merfeldaitė, O., Penkauskienė, D., Pivorienė, J., Railienė, A., Sadauskas, J., & Valavičienė, N. (2021). The value of critical thinking in higher education and the labor market: The voice of stakeholders. *Social Sciences*, 10(8), 286. <https://doi.org/10.3390/socsci10080286>
- Jöreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). *Multivariate analysis with LISREL*. Springer. <https://doi.org/10.1007/978-3-319-33153-9>
- Lai, K. (2018). Estimating standardized SEM parameters given nonnormal data and incorrect model: Methods and comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 600–620. <https://doi.org/10.1080/10705511.2017.1392248>
- Liang, X., & Yang, Y. (2013). Confirmatory factor analysis under violations of distributional and structural assumptions. *International Journal of Quantitative Research in Education*, 1(1), 61–84. <https://doi.org/10.1504/IJQRE.2013.055642>
- Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, 55(1), 102–119. <https://doi.org/10.1080/00273171.2019.1602503>

- Maturana-Moreno, G. A., & Lombo-Sánchez, M. L. (2020). Naturalistic intelligence: Effects on critical thinking and cognition needs. *Praxis & Saber*, 11(25), 177–204. <https://doi.org/10.19053/22160159.v11.n25.2020.9094>
- Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. CRC Press/Taylor & Francis Group. <https://doi.org/10.1201/9781439800393>
- Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: Segunda edición [Guidelines for test translation and adaptation: Second edition]. *Psicothema*, 25(2), 151–157. <https://doi.org/10.7334/psicothema2013.24>
- Nieto, A. M., Saiz, C., & Orgaz, B. (2009). Análisis de las propiedades psicométricas de la versión española del HCTAES-Test de Halpern para la evaluación del pensamiento crítico mediante situaciones cotidianas [Analysis of the psychometric properties of the Spanish version of Halpern's HCTAES-Test for assessing critical thinking using everyday situations]. *Revista Electrónica de Metodología Aplicada*, 14(1), 1–15. <https://doi.org/10.17811/rema.14.1.2009.1-15>
- Ortega-Quevedo, V., Gil-Puente, C., Vallés-Rapp, C., & López-Luengo, M. A. (2020). Diseño y validación de instrumentos de evaluación del pensamiento crítico en Educación Primaria [Design and validation of instruments for assessing critical thinking in primary education]. *Tecné, Episteme y Didaxis: TED*, 48, 91–110. <https://doi.org/10.17227/ted.num48-10557>
- Peters, G.-J. Y. (2018). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist*, 16(2), 56–69. <https://doi.org/10.31234/osf.io/h47fv>
- Rivas, S. F., Saiz, C., & Almeida, L. S. (2023). The role of critical thinking in predicting and improving academic performance. *Sustainability*, 15(2), 1527. <https://doi.org/10.3390/su15021527>
- Rodríguez, Á. M. M., González, A. E. J., & Velasco, N. Y. G. (2025). Fases para el diseño y validación de instrumentos de investigación: alcance de la inteligencia artificial. *Ciencia en Desarrollo*, (1), 105–109.
- Rodrigues, A., Soares, P., & da Silva, L. (2018). Translation, adaptation, and validation of the Halpern Critical Thinking Assessment to Portugal: Effect of disciplinary area and academic level on critical thinking. *Anales de Psicología*, 34(2), 292–298. <https://doi.org/10.6018/analesps.34.2.272401>
- Saiz, C., Almeida, L. S., & Rivas, S. F. (2022). Can critical thinking be assessed briefly? **Psico-USF, 26**(esp.), 139–148. <https://doi.org/10.1590/1413-8271202126nesp13>
- Saldarriaga-Zambrano, P. J., Gómez-Vergara, L. G., & Giraldo-Gutiérrez, J. A. (2022). Probabilidad e incertidumbre: Un aporte a la comprensión del pensamiento crítico en estudiantes universitarios [Probability and uncertainty: A contribution to the understanding of critical thinking in university students]. En M. D. García-Álvarez (Ed.), *Educational Revolution in the New Era* (Vol. 1, pp. 428–442). Redine.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Song, H., & Cai, L. (2024). Interactive learning environment as a source of critical thinking skills for college students. *BMC Medical Education*, 24(1), 270. <https://doi.org/10.1186/s12909-024-05247-y>
- Stensen, K., & Lydersen, S. (2022). Internal consistency: From alpha to omega. *Tidsskrift for Den Norske Lægeforening*, 142(12). <https://doi.org/10.4045/tidsskr.22.0288>

- Taseer, N. A., Siddique, A., Rabi, S. K., & Maqsood, A. (2023). Effect of emotional intelligence on students' academic performance in Pakistan. *Journal of Arts & Social Sciences*, *10*(1), 179–190.
- Tarasova, K., Gracheva, D., Talov, D., Orel, E., & Dementiev, A. (2025). Measuring changes in critical thinking skills among university economics students: Insights from domain-specific assessment. *Studies in Higher Education*, *50*(1), 1–16. <https://doi.org/10.1080/03075079.2023.2284000>
- Ventura-León, J., & Caycho-Rodríguez, T. (2017). El coeficiente omega: Un método alternativo para la estimación de la confiabilidad [The Omega coefficient: An alternative method for estimating reliability]. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, *15*(1), 625–627. <https://www.redalyc.org/pdf/773/77349627039.pdf>
- Wood, P. (2008). Confirmatory factor analysis for applied research. *The American Statistician*, *62*(1), 91–92. <https://doi.org/10.1198/tas.2008.s98>
- Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. <https://doi.org/10.3758/s13428-018-1055-2>

About the Contributor(s)

Nubia-Yaneth Gómez-Velasco, PhD in Education Sciences. Full Professor at Universidad Pedagógica y Tecnológica de Colombia. Tunja. Colombia. Member of the GAMMA Research Group and Hisula Group.

Email: nubia.gomez@uptc.edu.co

ORCID: <https://orcid.org/0000-0001-7745-1721>

Ana Emilce Jiménez-González, Msc. in Statistical Sciences. Associate Professor at Universidad Pedagógica y Tecnológica de Colombia. Tunja. Colombia. Member of the GAMMA Research Group.

Email: ana.jimenez@uptc.edu.co

ORCID: <https://orcid.org/0000-0003-0063-943X>

Myrian Esther Ortíz-Padilla, PhD in Education Sciences. Full Professor at Universidad Simón Bolívar. Barranquilla. Colombia. Member of the Educational and Social Synapse Research Group.

Email: myriam.ortiz@unisimon.edu.co

ORCID: <https://orcid.org/0000-0001-8964-9428>

Publisher's Note: *The opinions, statements, and data presented in all publications are solely those of the individual author(s) and contributors and do not reflect the views of Universitepark, EDUPIJ, and/or the editor(s). Universitepark, the Journal, and/or the editor(s) accept no responsibility for any harm or damage to persons or property arising from the use of ideas, methods, instructions, or products mentioned in the content.*
