Journal of Infrastructure, Policy and Development 2024, 8(15), 10089.
https://doi.org/10.24294/jipd10089

Article

Methodological proposal for analyzing critical thinking test scores as a tool for development in higher education

Nubia Yaneth Gómez-Velasco[1,*], Ana Emilce Jiménez-González[1], Myriam Esther Ortiz-Padilla[2]

[1] Universidad Pedagógica y Tecnológica de Colombia, Tunja 150001, Colombia
[2] Universidad Simón Bolívar, Barranquilla 080001, Colombia
*** Corresponding author:** Nubia Yaneth Gómez-Velasco, nubia.gomez@uptc.edu.co

CITATION

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

ARTICLE INFO

Received: 5 November 2024
Accepted: 2 December 2024
Available online: 9 December 2024

Abstract: The development of critical thinking (CT) enhances academic and professional opportunities. A review of literature reveals the use of fragmented analysis techniques, such as descriptive and correlational methods, among others, which hinder a deeper understanding of CT levels. This research aims to develop a methodology for analyzing Critical Thinking test scores, integrating five phases: exploratory, item analysis, scoring, gap analysis, and correlational. Using a quantitative approach, CT skills were analyzed with the Halpern Critical Thinking Assessment, which includes both open- and closed-ended questions to measure five skills: Verbal Reasoning (VR), Argument Analysis (AA), Hypothesis Testing (HT), Probability Use (PU), and Problem Solving (PS). The sample consisted of 214 students aged 18 and older. The item analysis phase categorized the items into quadrants: satisfactory, for review, or for elimination, based on difficulty and discrimination indices. The gap analysis revealed that Verbal Reasoning and open-ended formats were less satisfactory. The correlational phase, using heat maps, showed a stronger association between Verbal Reasoning and Probability Use. The methodological contributions include a variety of strategies that provide recommended procedures for analyzing tests or questionnaires in general. In today's digital age, the development of critical thinking is not only a desirable skill but an essential necessity for the higher education system.

Keywords: critical thinking; thinking skills; higher education; sustainability; assessment

1. Introduction

In today's digital era, where information flows freely and misinformation can skillfully masquerade as fact, critical thinking emerges not only as a desirable skill but as an essential necessity (Cioban and Mihaela, 2022; George, 2021). In this context, higher education institutions play a pivotal role in shaping citizens capable of discerning, analyzing, and evaluating information—skills closely linked to effective leadership processes and enhanced opportunities in their academic lives (Cangalaya, 2020; Gamboa et al., 2023). Motivated by this, it is imperative that educators, educational programs, and university entities actively commit to fostering critical thinking, as suggested by Acosta et al. (2020). This shared responsibility is crucial for equipping students with the necessary tools to tackle the complex challenges of today's world.

Critical Thinking (CT), according to Facione (2020), is broken down into interpretation, analysis, evaluation, and inferential processes, reflecting the richness and complexity of this concept from both methodological and contextual perspectives. Similarly, Elder and Paul (2005) emphasize CT as an intricate process of analysis and evaluation, highlighting the importance of a well-structured knowledge base and

creativity. Gutiérrez (2013) examines CT by focusing on the cognitive and argumentative components essential to its development. Additionally, Moreno-Pinado and Tejada (2017) propose methods to enrich CT by fostering skills in knowledge, comprehension, and introspection, offering a comprehensive view of the evolution and application of critical thinking in the educational field. Kaczkó and Ostendorf (2023) conceptualize CT as a validation of knowledge within the research model, with structures that promote problem-solving.

Based on Halpern's (2014) framework on critical thinking, five skills are assessed: Verbal Reasoning, Argument Analysis, Hypothesis Formulation, Problem Solving, and the Use of Probabilities. These skills are interconnected both within the teaching-learning process at different educational levels (Barra et al., 2021) and in research processes (Cabezas et al., 2017), as demanded by universities and society (Gómez et al., 2014; Gómez and Jiménez, 2015). Halpern's methodology remains relevant, as evidenced by recent studies evaluating Critical Thinking (CT) reported in the scientific literature by Acosta et al. (2020) and Lun et al. (2023). However, a persistent challenge in CT evaluation is the predominance of descriptive methodologies (Balatova et al., 2022; Gómez Velasco et al., 2023) and the limited use of statistical analyses (Sughra and Usmani, 2022), which results in a lack of methodological integration. This situation hinders the possibility of conducting more in-depth and comprehensive evaluations of CT.

In the current context of technological advances, higher education institutions face the challenge of adapting to unprecedented rates of change, which requires the implementation of educational strategies that promote critical thinking skills. The development of critical thinking is important for the professional growth of students, as well as for the progress and sustainability of the higher education system as a whole.

Within the described context, this research aims to propose an integrative five-phase methodology for analyzing Critical Thinking test scores: exploratory, item analysis, scoring, gap analysis, and correlational; taking into account sociodemographic variables such as academic semester, gender, and socioeconomic status. The proposal is enriched by incorporating a dual approach to assess open and closed-ended questions (Halpern, 2014), with a methodological framework that facilitates detailed analysis of Critical Thinking and can be adapted to other constructs and questionnaires, thus providing a versatile framework for educational research.

In this study, emphasis is placed on the notion that the analysis of instruments, such as those measuring Critical Thinking, can benefit from a methodology that integrates various techniques to obtain results capable of identifying item performance through psychometric analysis. Given the importance of having reliable and valid instruments, it is essential to explore analytical techniques that provide a more comprehensive evaluation, particularly in the context of Critical Thinking (CT).

The proposed methodology brings together a set of basic and advanced analytical techniques, including descriptive, correlational, and inferential analysis, as well as specialized indices for item evaluation, which allow for determining their difficulty and discrimination power. It incorporates correlation matrices visualized through heat maps and conducts gap analyses. This integrative approach facilitates a holistic and detailed understanding of the data, enabling a deeper and more precise evaluation of

Critical Thinking, which promotes the development of analytical and critical reasoning skills essential for professional and academic success.

2. Materials and methods

The research adopted a quantitative approach, facilitating both descriptive and inferential analyses focused on critical thinking competencies, utilizing Halpern's (2006) test as the assessment tool. The methodological design was characterized as exploratory, transactional, and correlational, thereby allowing for a comprehensive and multifaceted study of critical thinking skills within the evaluated context.

2.1. Sample

Through stratified random sampling, with a margin of error of 3% and a confidence level of 95%, a sample of 214 students was formed from the second (70), fifth (99), and eighth (45) semesters of the psychology program at Simón Bolívar University, located in northern Colombia. The average age of participants was 22 years (SD = 2.3). Participation by sex was distributed as 83% female, and by socioeconomic stratum, 84% belonged to strata 2 and 3, while 16% were from stratum 1.

2.2. Instrument

The measurement instrument used corresponds to the critical thinking test proposed by Halpern (2006), which includes both open- and closed-ended question formats for each of the five skills. The total score, as well as the score for each skill and format, is calculated according to the rubric provided in the manual (Halpern, 2006, 2016). A brief description is provided in **Table 1**.

Table 1. Description of skills and maximum possible scores, Halpern test (2016).

Aspect	Description	% of questions	Maximum score
Closed-ended question (multiple choice)	Critical thinking through recognition	49.0%	95
Open-ended question (essay)	Spontaneous critical thinking	51.0 %	99
Total critical thinking development score		100%	194
Critical thinking skills.			
Hypothesis Testing (HYPOTHESIS)	Analyzes a situation and identifies contradictions	25.8%	46
Verbal reasoning (verbal reasoning)	Facilitates comprehension and use of everyday language information.	8.8%	22
Argument analysis (argument analysis)	Identifies relevant information and beliefs	21%	41
Probability and uncertainty use (probabilities)	Applies logic to evaluate the likelihood of event	12.4%	24
Problem solving (problem solving)	Uses strategies to analyze and solve everyday problems	32.0%	61
Total		100	194

(Source: own).

This assessment has undergone validation in Spain and Chile for Spanish-speaking contexts. Specifically for Colombia, the research team leading this study has successfully advanced the validation process of the instrument, achieving a satisfactory outcome with a McDonald's ω reliability coefficient of 0.85.

2.3. Procedure and statistical analysis

Data collection was conducted in accordance with ethical principles, adhering to the regulations governing the handling of personal information. Participants were informed beforehand about the purpose of the research and voluntarily agreed to participate by signing an informed consent form. The data collection instrument was administered in a paper-and-pencil format. The dataset was compiled following a coding and quantification process of the scores obtained, categorized by skills and question formats. The statistical analysis was carried out across five phases, as outlined in **Table 2**.

Table 2. Methodological proposal for the score analysis of a test.

Methodological phase	Technique	Procedure
1. Exploratory phase (data preparation)	Data normalization	• Identification of the study unit. • Identification and classification of variables. • Valid scores for each variable according to Halpern (2006, 2016). • Review of redundancies and inconsistencies. • Excel file creation.
	Outlier detection (data entry errors, observed values)	• Identification of maximum and minimum scores for each variable and comparison with admissible values.
	Missing data analysis	• Review of responses provided by each study unit. • Establishment of a non-response threshold (in this study, a minimum of 10%) for excluding participants from analysis.
2. Item analysis	Difficulty index	• Proportion of correct responses. The closer the index is to zero, the more difficult the item. • Application of the formula using total, maximum, and minimum possible scores (Excel program)
	Discrimination index	• Difference in difficulty index between high and low scoring groups (Excel program).
	Discrimination coefficient	• Correlation between item score (variable) and corrected total score (total score excluding the item) (statistical programs: JASP, SPSS).
	Correlation quadrants	• Visual representation of difficulty and discrimination indices. Allows for item classification into satisfactory, improvable, or deficient (to be reviewed or removed). **Figure 1**. (SPSS program).
3. Score analysis		• Comparison of scores using descriptive measures. • Scatter plot (**Figure 2**). • Error bar chart (**Figure 2**).
4. Gap analysis	Comparison between observed (O) and theoretical (T) scores for each skill and question format.	• Error bar chart (**Figure 2**). • Percentage gap index (SPSS program).
5. Correlation analysis	Spearman correlation among skills	• Heatmap to determine the strength of relationships between scores (**Figure 2**) (Jamovi program).

(Source: own).

For greater clarity, some of the procedures outlined in **Table 2**, along with their evaluation criteria, are detailed below.

2.3.1. Phase: Normalization, outlier detection, and missing data

Before conducting the statistical analysis, it is essential to explore and clean the database. **Figure 1** outlines a five-step sequence (central arrows) where the applicable techniques are identified (upper boxes) and described (lower boxes).

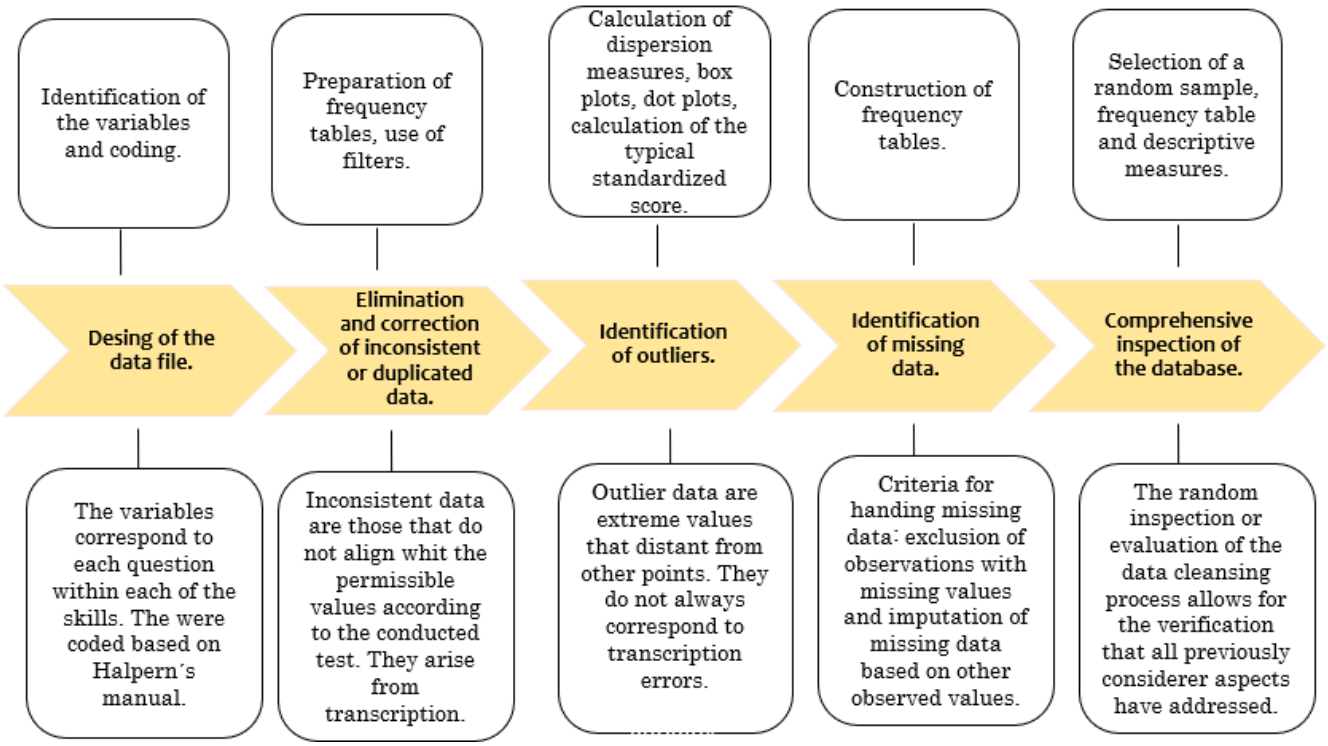


Figure 1. Sequence for normalization, identification of outliers and missing data.
(Source: own).

The exploration of the adequacy of the database provides organized and structured information to initiate the analysis process in accordance with the proposed objectives.

2.3.2 Item analysis phase

Difficulty index

The difficulty of an item determines how challenging or easy it is for a particular population, in terms of the participants who answer or solve it correctly. The difficulty index of an item can be determined by the proportion of individuals who answered it correctly (true or false items, multiple-choice), yielding values ranging from 0.0 to 1.0. However, for open-ended or graduated scoring items, which is the structure of Halpern's test, partially correct responses were considered, and the difficulty index was calculated using the formula proposed by Tjoe and de la Torre (2014): D = (fX-nX_min)/(n (X_max − X_min)).

Where fX is the total score obtained by all participants on an item, n is the number of participants, X_min is the smallest possible score for the item, and X_max is the highest possible score. Based on Adegoke's (2013) criterion, the items were classified as: acceptable (range of 0.3 to 0.7: moderate difficulty) or flawed (values below 0.3: difficult items, or values above 0.7: easy items).

Discrimination index

This index is defined as the ability of an item to differentiate between participants with high and low scores, based on those who answer correctly. To calculate it, the total scores were first obtained and ranked from highest to lowest, then organized into two groups following the Deborah, et al. (2020) criterion (the upper group comprising 27% of students with the highest scores, and the lower group comprising 27% of

students with the lowest scores). The discrimination index was calculated using the formula: ID = D_S − D_I.

Where D_S corresponds to the difficulty index of the upper group and D_I to that of the lower group. The values for this index range from −1 to 1. A negative value indicates that a greater number of low-scoring participants correctly answered the item, while a positive value suggests that a higher proportion of high-scoring participants answered it correctly. Following Hassan's (2016) criterion, an item is classified as having good discrimination when the index is greater than or equal to 0.40; moderately good when within the range of 0.30 to 0.39; fair for values between 0.20 and 0.29, indicating the item may need improvement; and poor for values less than or equal to 0.19, in which case the item should be reviewed or discarded.

Discrimination coefficient

This refers to the correlation between the corrected total score of the test and the score obtained on the item. Based on Beichner and Ding (2009), an item was considered to have good discrimination when its value exceeded 0.2.

Correlation quadrants

This construction is based on a scatter plot located in a Cartesian plane, where the *X*-axis represents the difficulty index values of the items with cut-off points at 0.3 and 0.7 (defined criteria for classification), and the *Y*-axis represents the discrimination index values with cut-off points at 0.2 and 0.3. The resulting graph is divided into nine quadrants, allowing for the classification of items into satisfactory, those with potential for improvement, and those that need to be reviewed or removed from the test. Additionally, to complement the analysis, the correlation between the difficulty and discrimination indices was obtained, which allowed for an assessment of the item's quality, according to Deborah et al. (2020).

2.3.3. Score analysis phase

Descriptive statistical measures of central tendency and dispersion were applied, following a review and identification of outliers. Outliers were identified using the interquartile range (IQR) method. Prior to conducting inferential tests, the normality of the distributions was assessed using the Shapiro-Wilk test, selected for its high power and effectiveness with moderate to small sample sizes. The results indicated significant deviations from normality ($p < 0.05$), which justified the use of the Mann-Whitney and Kruskal-Wallis tests to evaluate hypotheses regarding differences in average Critical Thinking (CT) scores based on skill, format, and sociodemographic variables such as gender, semester, and socioeconomic status. Additionally, homogeneity of variances was tested using Levene's test, which confirmed heteroscedasticity in several groups ($p < 0.05$), further supporting the use of non-parametric methods for analysis.

2.3.4. Gap analysis phase

The distance between the theoretical score and the actual score is referred to as the gap. To calculate it, the theoretical value T (maximum possible score according to Halpern) and the average value obtained by the students P were determined. Values closer to 100% indicate a larger gap, i.e., a greater distance between the obtained score

and the theoretical score. The percentage gap index was calculated using the following equation: $I_i = (T - P) / T \times 100$.

2.3.5. Correlation analysis phase

The hypothesis regarding the existence of a correlation between skills was tested using Spearman's Rho correlation coefficient, which is recommended for Likert scale questions (Hernández-Sampieri et al., 2018). A heat map was generated using the correlation matrix between skills. The color and its intensity represent levels of association between variables: red indicates an inverse correlation, white indicates no correlation, and green indicates a direct correlation. The greater the intensity of the color, the stronger the correlation.

This heatmap is a data visualization tool that uses colors to represent the magnitude of correlations between variables. The method is referred to as a "heatmap" because the color gradients resemble a thermal map, where shades vary according to the intensity of the represented phenomenon. This provides an intuitive and quick interpretation of the relationships between multiple variables at a glance.

For the organization of the database and calculations, Microsoft Excel was used, and for the statistical analysis, various software tools were employed: SPSS v.26 and JASP version 0.18.0.

3. Results and discussion

This section presents and analyzes the results in alignment with the proposed objectives and methodology, with a detailed breakdown of the findings across different sections.

3.1. Exploratory phase to adjust the database

Figure 2 presents the results of the adjustment on the five on the five steps described in **Figure 1**.

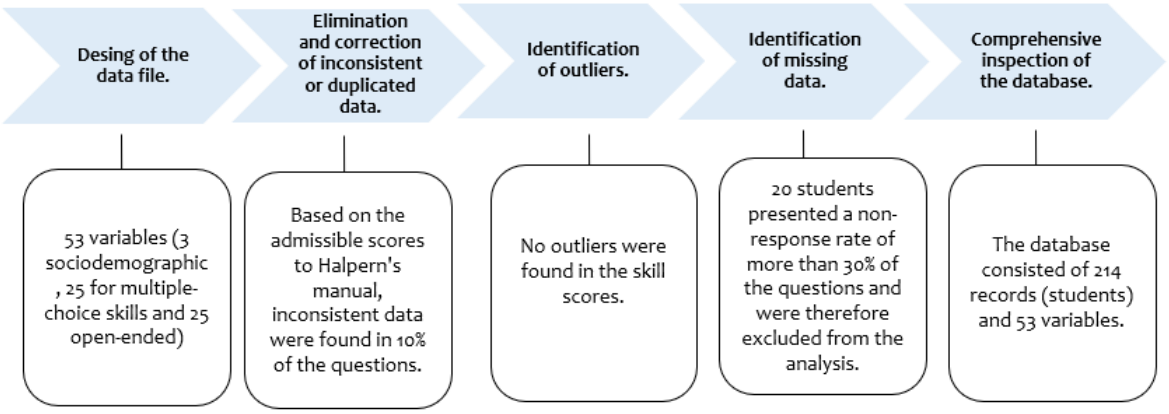


Figure 2. Exploratory phase results according to sequence.
(Source: own).

3.2. Item analysis

3.2.1. Test difficulty index

The difficulty index for all items ranged from 0.19 to 0.72, with an average of 0.42 (SD = 0.13), considered moderate (**Table 3**). More than half of the items (82%)

fell within the moderate range (0.3 to 0.7), while the remaining items were classified as difficult (16%) or easy (2%). For the closed-ended format, 44% of the items presented a difficulty index within the satisfactory range (0.3 to 0.58), whereas in the open-ended format, 38% of the items fell within this difficulty level (0.3 to 0.55).

All skills were found to be within the moderate difficulty range, with Argumentative Analysis being the most challenging and Problem Solving the least. Difficulty varied by academic semester, being higher for second-semester students, which suggests that students may develop PC skills progressively throughout their academic career. Regarding gender, the analysis shows that the difficulty index is similar for both men and women (**Table 3**).

3.2.2. Discrimination index

The discrimination index ranged from 0 to 0.47, with an average of 0.25 (SD = 0.12), indicating that the items are classified as having fair discrimination. Sixteen percent of the items showed good discriminatory power, 16% were moderately good, 28% were fair, and 40% were poor. However, when discrimination was measured using the item-total correlation, 18% exhibited high discriminatory power, 24% were moderately good, 38% were fair, and 20% were poor. In this case, 42% of the items were considered to contribute to the good internal consistency of the test.

Discrimination was higher for the items in the open-ended format compared to those in the closed-ended format (**Table 3**). Furthermore, Probability Use, Problem Solving, and Verbal Reasoning showed moderately good discrimination. Discriminatory power across semesters and between genders was found to be fair.

Table 3. Difficulty index, discrimination index, and discrimination coefficient by skill.

Type	**Difficulty**		**Discrimination**		**Correlation**	
Question	Media	SD	Media	SD	Media	SD
Open	0.37	0.09	0.34	0.12	0.34	0.11
Closed	0.47	0.14	0.22	0.12	0.27	0.15
Total	0.42	0.13	0.25	0.12	0.28	0.13
Skill						
Hypothesis	0.40	0.08	0.29	0.10	0.22	0.08
Verbal Reasoning	0.41	0.15	0.34	0.11	0.14	0.06
Argumentative Analysis	0.38	0.14	0.28	0.11	0.17	0.08
Probability Use	0.40	0.12	0.39	0.21	0.24	0.14
Problem Solving	0.53	0.09	0.35	0.12	0.45	0.07
Semester						
Second	0.39	0.14	0.24	0.16	0.25	0.18
Fifth	0.43	0.13	0.25	0.14	0.29	0.16
Eighth	0.46	0.15	0.22	0.16	0.23	0.19
Gender						
Women	0.43	0.13	0.24	0.12	0.27	0.12
Men	0.41	0.12	0.24	0.19	0.25	0.20

(Source: own).

3.2.3. Correlation quadrants. item classification into categories

Figure 3 relates the difficulty and discrimination indices by placing the items into 9 quadrants, identified by color according to the skill being assessed. Quadrants C1, C2, and C3 group the satisfactory items of the test, C4, C5, and C6 represent items with the potential for improvement, while the remaining three quadrants group deficient items that require revision or elimination.

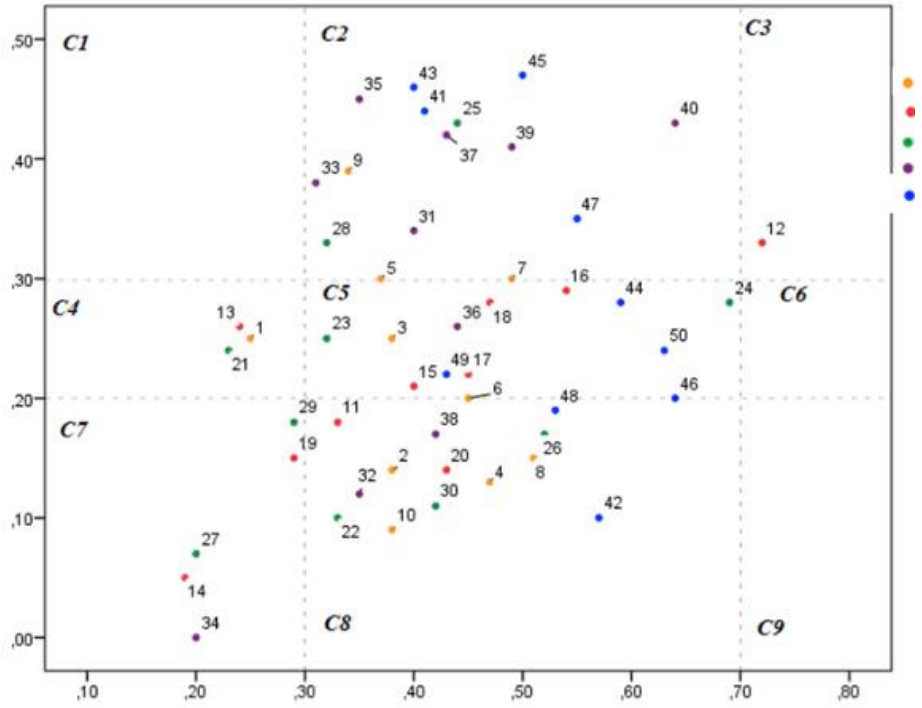


Figure 3. Classification quadrants based on difficulty index and discrimination index.
(Source: own).

Around 30% of the items fall within the 0.3 to 0.7 range for difficulty and above 0.3 for discrimination, indicating that these are satisfactory items (C2). Items located in C5 are potentially good, as they show moderate difficulty and discrimination with room for improvement. All skills contain satisfactory items, with a higher proportion found in Probability Usage and Problem Solving.

The correlation between the discrimination index and the difficulty index was significant (r = 0.28; P = 0.047), showing a direct relationship for the overall test. In the open-ended format, this correlation was significant and direct (r = 0.56; P = 0.04), as well as in the closed-ended format (r = 0.63; P = 0.01).

3.3. Score analysis

3.3.1 Comparisons by skill according to sociodemographic variables

Table 4 summarizes the five skills based on classifying variables and question formats. The averages and standard deviations are reported in parentheses, as well as the existence of significant differences between means at the 5% (**) and 10% (*) levels.

The values obtained reveal different performances. For instance, in Hypothesis Testing, there is a significant difference regarding academic semester, but not by sex or socioeconomic status. The average scores between the fifth and eighth semesters did not show significant differences.

Table 4. Critical thinking skills results. average score; standard deviation.

	Sex	**Semester**	**Socioeconomic Status**
Skills	(F;#178) (M;#32)	(2s;#70) (5s;#99) (8s;#45)	(E1;#27) (E2:#180)
1. Hypothesis testing (max theoretical 46 points)			
Total hypothesis	(19.2;5.0) (19.6;5.3)	(17.8;4.7)** (19.5;5.2) (20.4;5.2)	(18.4;6.1) (19.2;4.8)
Closed hypothesis	(19.2;5.0) (19.6;5.3)	(11.9;2.7) (11.9;3.3) (12.6;3.0)	(11.2;3.8) (12.2;2.9)
Open hypothesis	(7.1;3.4) (7.2;3.6)	(5.9;3.3)** (7.5;3.5) (7.8;3.5)	(7.2;3.4) (7.0;3.5)
2. Verbal reasoning (max theoretical 22 points)	Sex	Semester	Socioeconomic Status
Closed verbal reasoning	(3.3;1.3) (3.4;1.5)	(2.9;1.2)** (3.5;3.2) (3.5;1.2)	(3.2;1.4) (3.3;1.3)
Open verbal reasoning	(5.2;2.2) (5.1;2.1)	(4.8;1.9) (4.9;2.2) (6.4;2.3)**	(5.3;2.3) (5.2;2.2)
Total verbal reasoning	(8.5;2.9) (8.6;2.7)	(7.7;2.4) (8.4;2.9) (9.9;2.8)**	(8.6;3.1) (8.5;2.8)
3. Argumentation (max theoretical 41 points)	Sex	Semester	Socioeconomic Status
Closed argumentation	(9.7;2.8) (9.3;2.9)	(9.8;2.5) (9.2;3.2)* (10.1;2.6)*	(9.2;3.4) (9.7;2.7)
Open argumentation	(7.0;3.7) (6.8;3.4)	(6.9;3.7) (7.3;4.0)* (6.2;2.5)*	(5.7;2.8) (7.1;3.7)*
Total argumentation	(16.7;4.8) (16.1;4.9)	(16.7;4.7) (16.5;5.6) (16.3;3.3)	(14.9;5.2) (16.8;4.8)
4. Probability use (max theoretical 24 points)	Sex	Semester	Socioeconomic Status
Closed probability	(2.9;1.2) (3.0;1.2)	(2.6;1.2)** (2.9;1.2) (3.2;1.3)**	(3.2;1.2) (2.8;1.2)
Open probability	(7.0;3.8) (7.1;3.9)	(5.8;3.7)** (7.4;4.1) (7.6;3.4)	(7.2;4.3) (6.8;3.7)
Total Probability	(9.9;4.4) (10.0;4.6)	(8.5;4.3)** (10.3;4.5) (10.8;4.1)	(10.4;4.6) (9.7;4.3)

Table 4. (*Continued*).

	Sex	**Semester**	**Socioeconomic Status**
5. Problem Solving (Max Theoretical 61 points)	Sex	Semester	Socioeconomic Status
Closed Problem Solving	(23.7;4.9) (22.0;7.3)	(22.4;7.1)* (23.3;5.4) (24.5;4.0)*	(22.9;6.4) (23.8;4.9)
Open Problem Solving	(10.4;4.4) (8.8;4.8)	(9.5;5.2) (9.4;4.1) (12.3;4.0)**	(9.6;5.1) (10.3;4.4)
Total Problem Solving	(34.1;7.6) (31.1;10.8)*	(31.9;10.3) (32.7;8.3) (36.8;6.8)**	(32.5;10.1) (34.1:7.6)
Totals			
Total Closed (Max Theoretical 95 points)	(51.7;9.4) (50.3;10.6)	(49.6;9.6) (50.9;10.7) (54.0;7.8)**	(49.7;13.8) (51.8;8.7)
Total Open (Max Theoretical 99 points)	(36.7;7.3) (35.1;12.0)	(33;13.6)* (36.5;13.8) (40.4;10.3)*	(34.9;14.5) (36.7;12.9)
Total Skills (Max Theoretical 194 points)	(88.4;19.0) (85.4;19.2)	(82.6;19.1) (87.3;20.7) (94.3;16.0)*	(84.7;25.9) (88.4;17.7)

(Source: own).
Note: Maximum(max.), Reasoning(reason.), Resolution(res.), Skills(skill.).

In most skills, there were no significant differences based on gender, except in total Problem Solving (the sum of closed and open formats), where women scored on average 3 points higher.

When comparing socioeconomic statuses 1 and 2, a statistically significant difference was found in Argument Analysis (open format), with higher averages observed among students from socioeconomic status 2. The difference was approximately 1.5 points, indicating greater ability in Argument Analysis among students from higher socioeconomic backgrounds.

All critical thinking skills showed statistically significant differences by semester in either one of the formats (open or closed), or in both. The averages increased as students progressed through their academic semesters, particularly when comparing second-semester students with eighth-semester students. Significant differences in some skills, such as Verbal Reasoning, were observed between fifth and eighth-semester students (**Table 4**).

3.3.2. Score comparisons for open and closed-format questions

The scatter plot (**Figure 4a**) illustrates the behavior of scores obtained from open and closed-format questions by semester. There is a higher concentration of mid to high scores in both open and closed-format questions for eighth-semester students, in contrast to the scores of fifth and second-semester students.

Figure 4b presents confidence interval bars for the total average scores, derived from prior data standardization due to the differences in score ranges between open and closed formats, according to the Halpern test. As students' progress through their semesters, there is a slight increase in scores for both formats. Second-semester

students show slightly better performance in the closed format, which is contrary to the trend observed in fifth and eighth-semester students.

3.4. Gap analysis: Theoretical vs. observed performance

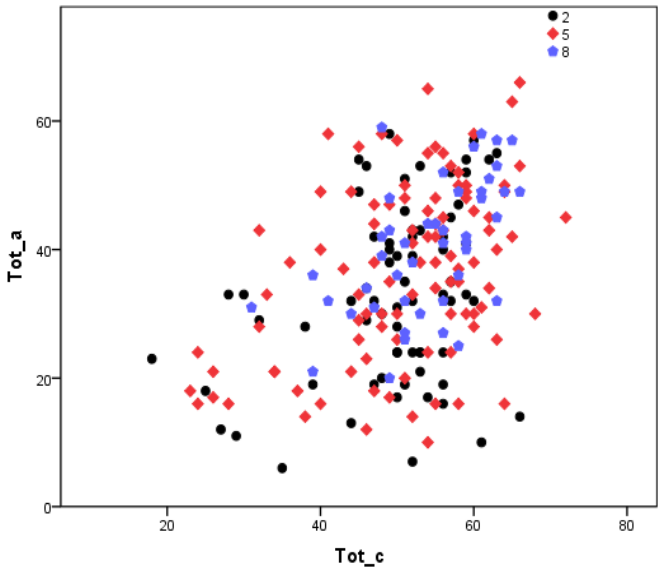
Average scores obtained in different skills are compared against the theoretical maximum possible (Halpern, 2016). All skills fall below the theoretical average, except for Problem Solving, which implies lower-than-expected performance in four of the five skills (**Figure 4c**).

The skills with the largest gaps are Verbal Reasoning, Argument Analysis, Use of Probability, and Hypothesis Testing, with respective gap indices of 61.4%, 59.7%, 59.2%, and 58.4%. The smallest gap was found in Problem Solving, with a gap of 45.4%. The maximum theoretical score for the open format was 99.0 points, and the average score obtained by students was 36.2, indicating a gap of 62.9 points and an index of 63.4%. For the closed format, the gap was 43.9 points, with a gap index of 46.2%. Based on these results, it can be inferred that students generally performed better on closed-format questions.

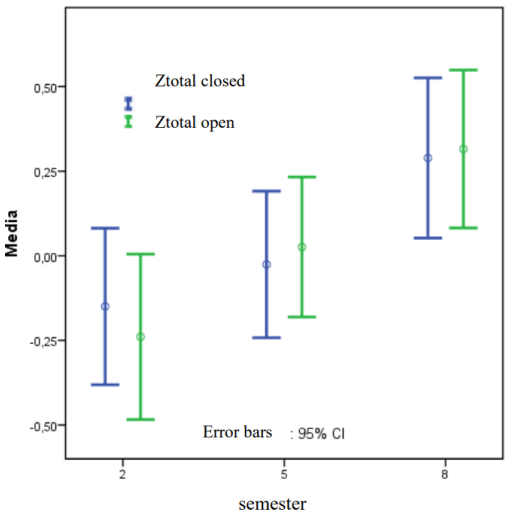
3.5. Correlation analysis with heatmap diagram

The correlation matrix, utilizing Spearman's Rho coefficient (**Figure 4d**), identifies the presence of a moderate and significant correlation among all skills, visualized according to color intensity. Notably, there exists a stronger moderate and significant correlation between Verbal Reasoning and Probability Use (0.529). Additionally, a moderate correlation is observed between the total scores of open-ended and closed-ended question formats (0.398).

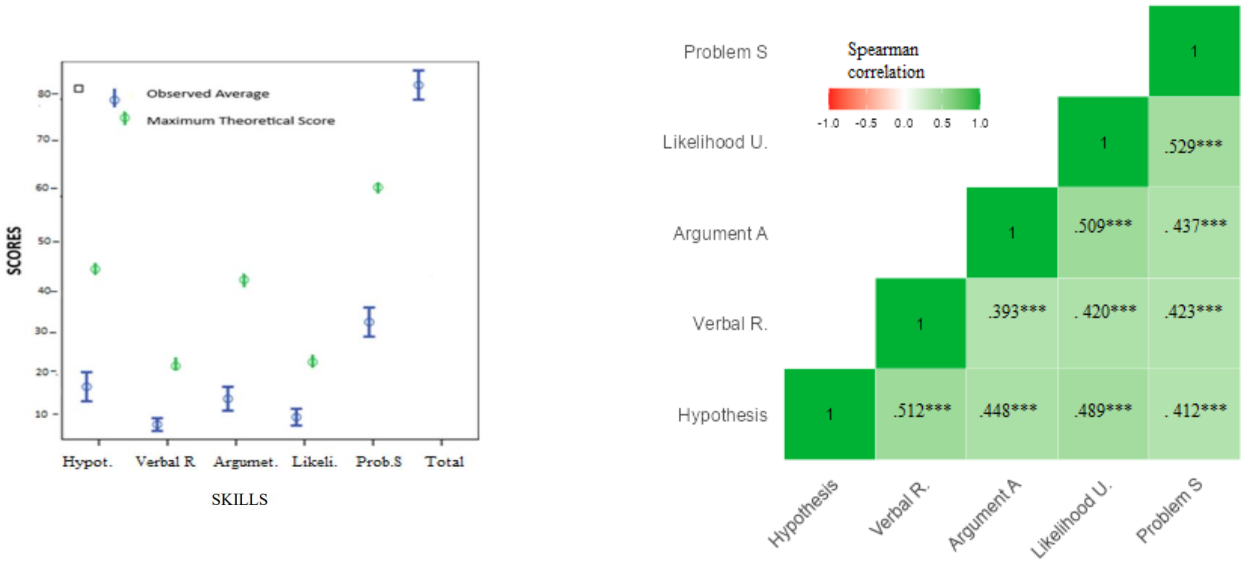
The correlation analysis between skills established that all of them contribute to measuring Critical Thinking (CT) and are not redundant, as their values are not excessively high. Moreover, the correlation serves as an indicator of the test's multidimensional structure and the potential independence between factors.



(a)



(b)



(c) **(d)**

Figure 4. Score Comparisons in open and closed question formats.
(Source: own).

4. Discussion

The study reveals decisive procedures across five phases for the effective assessment of Critical Thinking (CT) development.

The exploratory phase was crucial for guiding the cleaning and structuring of data, enabling a more efficient analysis that facilitated the identification of outliers and transcription errors, both critical elements for improving the study's integrity. The subsequent item analysis phase provided a more detailed evaluation of item quality, focusing on difficulty and discrimination, which are essential aspects for tests of this nature, as suggested by Rivas and Saiz (2012). Comparatively, studies like that of Bertea and Zait (2013) have reported similar findings regarding the usefulness of these preliminary phases in improving the quality of measurement instruments. However, other studies, such as Ehido et al. (2020), have highlighted limitations in the ability of these phases to detect certain types of errors in diverse educational contexts, emphasizing the importance of additional methodological adjustments to ensure the generalizability of results.

In the score and correlation analysis phases, statistical techniques were applied to test hypotheses related to the test scores, based on sociodemographic variables. The gap analysis highlighted CT skills requiring greater attention, which, once identified, prompt actions from educators and institutions aimed at developing multidimensional strategies for improvement (Christensen et al., 2023), including educational technologies (Elreda and Kohler, 2023).

The evaluation of the scores revealed a tendency toward higher CT scores among women compared to men, with an average difference of three points, although this was not statistically significant. This finding aligns with Sughray and Usmani (2022), who also observed similar patterns with minor contextual variations. Additionally, an increase in average scores was noted as students progressed in their university education, with advanced-semester students outperforming those in earlier semesters, a trend consistent with the findings of Acosta et al. (2020). This progression likely

reflects greater exposure and skill in handling critical competencies, underscoring the need for early educational interventions to strengthen these abilities from the initial years of university studies.

The correlation between skills indicated significant direct associations, particularly between Verbal Reasoning and the Use of Probabilities, which was highlighted by a more intense color in the heatmap. These correlations can help identify key associations and guide targeted improvement efforts.

This article enriches the field of Critical Thinking (CT) by offering a variety of strategies and a series of detailed procedures in each phase. It facilitates the use of a CT assessment instrument that, through its quantitative, empirical, and standardized nature, provides evidence of this essential competency, as suggested by Manassero and Vázquez (2020). Furthermore, this work broadens the spectrum of evaluations and analyses by complementing traditional qualitative approaches, as discussed in the research of Costa et al. (2021) and Aji et al. (2023), thereby contributing to a comprehensive view of CT measurement.

A notable aspect of this study is the analysis of a test type that not only incorporates closed-format questions—common in most evaluations, as noted by Rivas and Saiz (2012)—but also includes open-ended questions based on real-life situations, as reported by Díaz-Larenas et al. (2017).

This dual approach highlights the importance of preparing educators to properly analyze these tests, which are key tools for diagnosing and evaluating the development of CT in students, as emphasized by Barreiro et al. (2021), Balatova et al. (2022), and Covaleda et al. (2023).

5. Conclusion

The research presents an innovative methodology for analyzing the scores of a test that evaluates Critical Thinking (CT) in university students, integrating both advanced and fundamental analytical techniques across five distinct phases.

The proposed methodology contributes significantly to the evaluation of CT in university students by integrating both classical and advanced analytical techniques.

The implementation of the five developed phases enables educators and universities to obtain results that strengthen educational methods, adapt academic programs, and prepare graduates to face real-world challenges. The research emphasizes the effective and coordinated use of methodological phases, encouraging their application to evaluate and analyze the development of CT in university students. However, it also suggests extending their application to other educational levels and adapting them to assess other constructs.

Although a methodology has been proposed that integrates analysis techniques little reported in studies of tests such as the one that measures CT, for future studies it is proposed to incorporate other techniques to analyze the construct with variables according to the context of the participants. Combining qualitative and quantitative techniques to give greater support to the analysis and thus be able to draw conclusions that allow the creation of strategies for the strengthening of competencies for CT

Limitations: The application of this methodology may face limitations related to the theoretical foundation required by the researcher at each phase, as well as the

handling of various software programs, since not all techniques are implemented within a single platform. Additionally, the sample selection requires proportional sizing for each of the derived groups, as this could affect the generalization of the results and their applicability in diverse educational contexts. These limitations present a challenge in terms of accessibility and ease of implementation for some researchers and educators.

Future Research Lines: Future research lines propose the inclusion of techniques that allow for the joint analysis of multiple characteristics, such as cluster analysis, correspondence analysis, regression, and multiple correlation, among others. This would enable a deeper understanding of the various characteristics of CT and its development among university students. It is suggested to explore the effectiveness of the methodological framework in different educational contexts, in order to evaluate the application and effectiveness of the proposed methodology. Longitudinal studies could also be implemented to monitor the development of CT in students throughout their university education, with the goal of identifying more effective interventions.

Knowledge transfer: Three key areas of knowledge transfer resulting from this research are highlighted.

It enables universities to assess and enhance students' critical thinking skills. By identifying and strengthening these essential abilities, it contributes to preparing young individuals for innovation and entrepreneurship processes that can drive the economic growth of the region and the country.

From a social and educational perspective, the adoption of this methodology has a social impact by providing educators and institutions with evaluation strategies for the development of CT in their students. This allows for results that identify areas for improvement, helping to shape professionals capable of effectively addressing social issues and actively participating in evidence-based decision-making.

Regarding academic innovation, the article provides a solid foundation for academic research and development, offering multiple articulated tools to assess and improve critical thinking. The findings and methodologies can be applied across various fields of study and constructs, facilitating interdisciplinarity and promoting the evaluation and analysis of results in higher education and other educational levels.

In this context, the proposed methodological framework, by establishing an innovation that combines and integrates both fundamental and advanced techniques across its five phases, raises the standard of CT evaluation in higher education as well as in other educational levels. Its applicability extends further, allowing the evaluation of various constructs, demonstrating its versatility and potential to enrich the educational process across a wide range of contexts.

The adoption of methodologies for the analysis of critical thinking test scores contributes to the strategic development of the higher education system. This study evidences the need to integrate evaluative approaches that improve the diagnostic capacity of critical competencies among students, and in turn contribute to strengthen the foundations of an educational model for innovation and social progress. By fostering an academic environment conducive to critical thinking, educational institutions become agents of change, preparing students to face the challenges of the present and lead the sustainable and comprehensive development of the future.

Author contributions: Conceptualization, MEOP and NYGV; methodology, AEJG and NYGV; software, AEJG and NYGV; validation, MEOP, AEJG and NYGV; formal analysis, AEJG and NYGV; investigation, MEOP, AEJG and NYGV; resources, NYGV and AEJG; data curation, MEOP; writing—original draft preparation, MEOP, AEJG and NYGV; writing—review and editing, AEJG and NYGV; visualization, MEOP; supervision, MEOP and AEJG; project administration, NYGV and AEJG funding acquisition, AEJG and NYGV. All authors have read and agreed to the published version of the manuscript.

Funding: The authors declare that they received financial support for the research, authorship, and/or publication of this article from: Universidad Pedagógica y Tecnológica de Colombia and Universidad Simón Bolívar de Colombia- SGI 3716.

Patient or public contribution: The study involved voluntary and de-identified participation of community college students in the survey.

Data availability statement: The data utilized in this study are not publicly accessible initially, as their dissemination requires prior authorization from the funding institutions. Any request for access to the data must be approved by the respective institutions.

Conflict of interest: The authors declare no conflict of interest.

References

Acosta Meza, D., Atencia Andrade, A., García Medina, M. A. & Rodríguez Sandoval, M. (2020)) 'Identificación del pensamiento crítico en estudiantes universitarios de segundo semestre de la Corporación Universitaria del Caribe (CECAR)' (Spanish) , Revista Electrónica Interuniversitaria de Formación del Profesorado, 23(3). doi: https://doi.org/10.6018/reifop.435831.

Adegoke, B. A. (2013) 'Comparison of item statistics of physics achievement test using classical test and item response theory frameworks', Journal of Education and Practice, 4(22), pp. 87-96. Available at: https://www.semanticscholar.org/paper/Comparison-of-Item-Statistics-of-Physics-Test-using-Adegoke/731a206284fad1d2522b02337b6c5409a8e91bb5 (Accessed: 17 July 2023).

Aji, M. P., Negoro, R. A., Rusilowati, A. & Subali, B. (2023) 'Development of Waves Critical Thinking Test: Physics Essay Test for High School Students', Development Of Waves Critical Thinking Test: Physics Essay Test For High School Students. Available at: https://www.eu-jer.com/development-of-waves-critical-thinking-test-physics-essay-test-for-high-school-student (Accessed: 17 July 2023).

Balatova, K., Nepras, K., Kovarik, P., Kubiatko, M. & Sustekova, E. (2022) 'The Influence of Selected Variables on University Students' Critical Thinking Level: Preliminary Results'. doi: 10.26907/esd.17.4.03.

Barra, M., Lopez, M., Moreno, E. & Uyaguari, F. (2021) 'Developing critical thinking in the classroom: testimonies from excellent Ecuadorian teachers'(Spanish), Areté. Digital Journal of the Doctorate in Education of the Central University of Venezuela, 8(15), pp. 161–180. doi: https://doi.org/10.55560/ARETE.2022.15.8.8.

Barreiro, M. P. R., Barreiro, J. R., Bravo, K. L. M., Colamarco, I. L., Velásquez, B. I. H. & Rivadeneira, L. (2021) 'Critical thinking and its assessment in university education'(Spanish), Research, Society and Development, 10(3), pp. e51910313748-e51910313748. doi: https://doi.org/10.33448/rsd-v10i3.13748.

Beichner, R. & Ding, L. (2009) 'Approaches to data analysis of multiple-choice questions', Physical Review Special Topics-Physics Education Research, 5(2), 020103. doi: https://doi.org/10.1103/PhysRevSTPER.5.020103.

Bertea, E., & Zait, A. (2013). Scale Validity In Exploratory Stages Of Research. Management and Marketing Journal, 38-46.

Cangalaya, L. (2020) 'Critical thinking skills in university students through research'(Spanish), From the South Magazine,12(1), pp. 141-153. doi: https://doi.org/10.21142/DES-1201-2020-0009.

Cabezas, J. A. A., Cepeda, M. P. & Cordova, R. M. Z. (2017) 'Critical thinking applied to scientific research' (Spanish), Atlante Magazine: Education and Development Notebooks. Available at: http://www.eumed.net/rev/atlante/2017/02/investigacion.html (Accessed: 17 July 2023).

Christensen, R., Knezek, G., Smits, A., Tondeur, J. & Voogt, J. (2023) 'Strategies for developing digital competencies in teachers: Towards a multidimensional Synthesis of Qualitative Data (SQD) survey instrument', Computers & Education, 193, 104674. doi: https://doi.org/10.1016/j.compedu.2022.104674.

Cioban, M. & Mihaela, H. (2022) 'Impactul tehnologiilor educaționale moderne în formarea competenței matematice la elevii din învățământul profesional tehnic postsecundar nonterțiar (viitori învățători)', Universitatea de stat din Tiraspol universitatea pedagogică de stat „Ion Creangă". Available at: http://www.cnaa.md/files/theses/2022/58564/mihaela_hajdeu_thesis.pdf (Accessed: 17 July 2023).

Costa, C., Hart, C., D'Souza, D., Kimpton, A. & Ljbusic, J. (2021) 'Exploring higher education students' critical thinking skills through content analysis', Thinking Skills And Creativity, 41, 100877. doi: https://doi.org/10.1016/j.tsc.2021.100877.

Covaleda, I., Fayos, L., Murillo-Ligorred, V. & Ramos-Vallecillo, N. (2023) 'Knowledge, Integration and Scope of Deepfakes in Arts Education: The Development of Critical Thinking in Postgraduate Students in Primary Education and Master's Degree in Secondary Education', Education Sciences, 13(11), 1073. doi: https://doi.org/10.3390/educsci13111073.

Deborah, A. C., Peter, A. O. & Temitope, B. (2020) 'Comparism of Item Difficulty and Discrimination of Pre And Post University Entrance Examinations in Nigeria', The Universal Academic Research Journal, 3(1), pp. 1-9. Available at: https://dergipark.org.tr/en/pub/tuara/issue/62346/937932 (Accessed: 17 July 2023).

Díaz-Larenas, C. H., Martín, L. S., Nelly, G., Ossa-Cornejo, C. J., Palma-Luengo, M. R. & Quintana-Abello, I. M. (2017) 'Analysis of instruments for measuring critical thinking', Psychological Sciences, 11(1), pp. 19-28. doi: https://doi.org/10.22235/cp.v11i2.1343.

Ehido, A., Awang, Z., Halim, B., & Ibeabuchi, C. (2020). DEVELOPING ITEMS FOR MEASURING QUALITY OF WORK LIFE AMONG MALAYSIAN ACADEMICS: AN EXPLORATORY FACTOR ANALYSIS PROCEDURE. Humanities & Social Sciences Reviews. https://doi.org/10.18510/hssr.2020.83132

Elder, L. & Paul, R. (2005) Critical Thinking Competency Standards: Standards, principles, performance, indicators, and outcomes, with a master rubric on critical thinking.(Spanish) Available at: https://www.criticalthinking.org (Accessed: 17 July 2023).

Elreda, L. M. & Kohler, E. A. (2023) 'EdTech Context Inventory: Factor analyses for ten instruments to measure edtech implementation context features', Computers & Education, 195, 104709. doi: https://doi.org/10.1016/j.compedu.2022.104709.

Facione, P. A. (2011) Critical thinking: What it is and why it counts, Insight Assessment. Available at: https://insightassessment.com/wp-content/uploads/ia/pdf/whatwhy2018.pdf (Accessed: 17 July 2023).

Gamboa Suárez, A. A., Montes Miranda, A. & Rosales Yepes, A. (2023) 'Effective leadership for educational quality and university accreditation in the Colombian Caribbean'(Spanish), Interuniversity Journal of Teacher Training, 98(37.1). doi: https://doi.org/10.47553/rifop.v98i37.1.98213.

George Reyes, C. E. & Glasserman Morales, L. D. (2021) 'Technology-mediated research competencies: a systematic mapping of the literature' (Spanish), Education in the Knowledge Society (EKS), 22, e23897. doi: https://doi.org/10.14201/eks.23897.

Gómez Velasco, N. Y., Jiménez González, A. E., & Rodríguez Gutiérrez, J. K. (Eds.). (2023). Elementos de bibliometría. Fundamentos y aplicaciones. Publicaciones Universidad de América. https://doi.org/10.29097/9786289517651 (Original work published 11 de abril de 2023)

Gómez, N. Y. & Jiménez, G. A. E. (2015) 'Statistics as a support for university-community research projects. Reflections on an experience with research incubators' (Spanish), Logos, Science & Technology Magazine, 7(1), pp. 27-34. doi: http://dx.doi.org/10.22335/rlct.v7i1.210.

Gómez, N. Y., Jiménez, A. E., Guerrero, S. C. & Ayala, Y. (2014) 'Analysis of Colombian scientific production in chemistry. WoK database (2001-2012)' (Spanish), Logos Ciencia & Tecnología Journal, 6(1), pp. 108-115.

Gutiérrez, C. F. V. (2013) 'An interdisciplinary reflection on critical thinking', Latin American Journal of Educational Studies (Spanish),9(2), pp. 11-39. Available at: https://www.redalyc.org/articulo.oa?id=134135724002 (Accessed: 17 July 2023).

Halpern, D. F. (2006) Halpern critical thinking assessment using everyday situations: Background and scoring standards (2nd Report), Claremont McKenna College. Available at:

http://www.scielo.org.co/scielo.php?script=sci_nlinks&ref=000116&pid=S0121-38142014000200006000015&lng=en (Accessed: 17 July 2023).

Halpern, D. F. (2014) Thought and knowledge: An introduction to critical thinking (5th ed.), Psychology Press.

Halpern, D. F. (2016) Manual Halpern Critical Thinking Assessment, Schuhfried GmbH, Mödling.

Hassan,S. (2016) 'item analysis, reliability statistics and standard error of measurement to improve the quality and impact of multiple-choice questions in undergraduate medical education in faculty of medicin at UNISZA', Malaysian Journal of Public Health Medicine, pp.7-15. Available at: http://wprim.whocc.org.cn/admin/article/articleDetail?WPRIMID=626840&articleId=626840 (Accessed: 17 July 2023).

Hernández Sampieri, R., Fernández Collado, C. & Baptista Lucio, P. (2018) Research methodology (4th ed.), México: McGraw-Hill Interamericana.

Kaczkó, É. & Ostendorf, A. (2023) 'Critical thinking in the community of inquiry framework: An analysis of the theoretical model and cognitive presence coding schemes', Computers & Education, 193, 104662. doi: https://doi.org/10.1016/j.compedu.2022.104662.

Lun, V. M. C., Yeung, J. C. & Ku, K. Y. L. (2023) 'Effects of mood on critical thinking', Thinking Skills and Creativity, 47, 101247. doi: https://doi.org/10.1016/j.tsc.2023.101247.

Manassero-Mas, M. A. & Vázquez-Alonso, Á. (2020) 'Assessing Critical Thinking Skills: Validation of Culture-Free Instruments' (Spanish), Tecné, Episteme y Didaxis: TED, (47), pp. 15-32. doi: https://doi.org/10.17227/ted.num47-9801.

Moreno-Pinado, W. E. & Tejeda, M. E. V. (2017) 'Teaching strategy to develop critical thinking'(Spanish) , REICE. Ibero-American Journal on Quality, Effectiveness and Change in Education, 15(2), pp. 53-73. Available at: https://www.redalyc.org/articulo.oa?id=55150357003 (Accessed: 17 July 2023).

Rivas, S. F. & Saiz, C. (2012) 'Validation and psychometric properties of the PENCRISAL critical thinking test' (Spanish), REMA Electronic Journal of Applied Methodology, 17(1), pp. 18-34. Available at: http://hdl.handle.net/10366/157532 (Accessed: 17 July 2023).

Severín, E. (2011) Technologies for Education (TEd). A Framework for Action, Inter-American Development Bank.

Sughra, U. & Usmani, A. (2022) 'Comparison of Critical Thinking among undergraduate medical students of Conventional and Integrated curricula in Twin Cities', Pakistan Journal of Medical Sciences, 38(6), pp. 1453. doi: 10.12669/pjms.38.6.5409.

Tjoe, H. & de la Torre, J. (2014) 'The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework', Mathematics Education Research Journal, 26