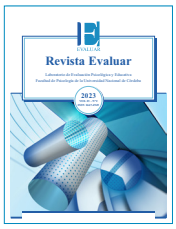
2023, Vol. 23, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



Assessing Critical Thinking Skills: A Diagnosis of Elementary Students

Evaluación de las destrezas del pensamiento crítico: Un diagnóstico de los estudiantes de primaria

María Antonia Manassero-Mas [1] , Ángel Vázquez-Alonso * [2]

1 - Facultad de Psicología, Universidad de las Islas Baleares (España).
2 - Instituto de Investigación e Innovación Educativa, Universidad de las Islas Baleares (España).

Introduction
Method
Results
Discussion
References

Recibido: 28/02/2023 **Revisado:** 05/03/2023 **Aceptado:** 16/03/2023

Abstract

Critical thinking (CT) is a central aim of the 21st century education, although its research lacks consensus, has been unequal and lacking in primary evaluation issues and primary education. These weaknesses justify the aim of this study: evaluating primary students' mastery of CT skills. The quantitative methodology diagnoses six CT skills (prediction, logical reasoning, comparison, classification, decision-making, and problem-solving) through a 48-item test, which 655 sixth-graders completed. The students display an average global mastery of the CT items and skills, and the test's and skill scores' standardization are also presented. The comparison of boys and girls shows that girls perform better than boys on most test items. The diagnoses suggest an intermediate mastery of CT skills in students, where girls widely outperform boys, and also propose some educational implications.

Keywords: *assessment, critical thinking, skills, normative data in primary education, gender differences*

Resumen

El pensamiento crítico (PC) es un objetivo central de la educación del siglo XXI, aunque su investigación carece de consenso y ha sido desigual, y limitada en temas de evaluación en la educación primaria. Estas debilidades justifican el objetivo de este estudio: evaluar el dominio de las destrezas de PC en estudiantes de primaria. La metodología cuantitativa evalúa seis destrezas de CT (predicción, razonamiento lógico, comparación, clasificación, toma de decisiones y resolución de problemas) a través de un test de 48 ítems (alfa = .85), en el que participaron 655 estudiantes de sexto grado (11 años). Los estudiantes exhiben un dominio general promedio de los ítems y destrezas del pensamiento crítico. Además, se presentan las estandarizaciones de los resultados de los test y el puntaje obtenido en los ítems. La comparación de PC entre niños y niñas indica que las niñas obtienen mejores puntuaciones que los niños en la mayoría de los ítems. Con el estudio se concluye que existe un dominio intermedio del PC, que es ampliamente mejor en niñas, y se discuten algunas implicaciones educativas.

Palabras clave: *evaluación, pensamiento crítico, destrezas, baremación en primaria, diferencias de género*

***Correspond to:** Ángel Vázquez-Alonso. Dirección: Universidad de las Islas Baleares, Edificio Guillem Cifre de Colonya, Carretera de Valldemossa, km. 7.5, 07122, Palma, España. Teléfono: +34 971 17 30 75; FAX: 34 971 17 31 90. E-mail: angel.vazquez@uib.es
Author´s Note: Proyecto EDU2015-64642-R (AEI/FEDER, UE) financiado por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER).
How to cite: Manassero-Mas, A. M., Vázquez-Alonso, A. Assessing Critical Thinking Skills: A Diagnosis of Elementary Students (CT). *Revista Evaluar, 23*(2), 40-56. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar
Participaron en la edición de este artículo: Agustina Mangieri, Gloria Rebeca Nieve, Juan Cruz Balverdi Nieto, Rita Hoyos, Florencia Ruiz, Jorge Bruera.

Introduction

Many institutions and experts worldwide support educating students for the skills of the 21st century to face the great challenges of today (European Union, 2014; Fullan & Scott, 2014; International Society for Technology Education, 2003; National Education Association, 2012; National Research Council, 2012; OECD, 2018; UNESCO, 2016). These skills include digital and cognitive skills; the latter usually distinguishes soft (psychosocial or interpersonal) and hard (higher-order cognitive) skills, which some authors summarize in the 4Cs or 6Cs (collaboration, communication, character, citizenship, creativity, and critical thinking [CT]). In sum, CT is a significant component of the skills for the 21st century, placing innovative demands on education (Almerich et al., 2020; Vincent-Lancrin et al., 2019).

From an educational perspective, CT teaching aligns with Piaget's pioneering studies (Piaget & Inhelder, 1997) and cognitive acceleration programs (Shayer & Adey, 2002) that have empirically demonstrated its significant impact on learning. In addition, the cognitive skills that make up CT are connected to the higher categories of Bloom's taxonomy (analyze, judge, and create), they are often called higher-order thinking skills. However, they also require the most basic skills, knowledge and understanding (Krathwohl, 2002). Nowadays, the mastery of CT skills is considered a key factor in achieving meaningful and deep learning skills that characterize educational excellence (Valenzuela, 2008). The meta-analysis of Hattie reports that the effect size of Piagetian programs on learning is very large ($d = 1.28$), and the impact of different CT skills (metacognitive strategies, creativity, problem-solving, etc.) is also high ($d > .40$) (Hattie, 2009).

From a labor perspective, most surveys show that CT is a primary and invariable requirement of future jobs (World Economic Forum, 2021) and a key factor for people's success in the information age (Tremblay et al., 2012). This labor requirement, coupled with the evolution of cognitive development, have driven most of the innovative teaching efforts of CT to be focused on higher education.

In sum, CT is a central objective of education, an important attribute of citizenship in a democratic society, and a decisive factor of an individual's professional success in the 21st century. These beneficial characteristics justify the attention placed on CT as a central variable of school learning. This study approaches this idea from a diagnostic evaluation perspective to address the lack of information about younger students' CT skills and aims to present this information from primary education and thus contribute to fill this gap.

Critical thinking

Research on CT has focused on 3 areas: conceptualization, teaching, and evaluation. However, the development of each area has been unequal (Saiz, 2017).

In the framework of cognitive psychology, CT is generally conceptualized as a type of thinking that masters multiple higher-order cognitive skills and various attitudinal dispositions, and is regulated by demanding quality standards (precision, solidity, coherence, relevance, adequacy, etc.) to overcome thinking's natural tendencies toward error, fallacy, and bias (egocentrism and socio-centrism). These skills, provisions, rules, and values inherent in CT provide a crucial basis for its evaluation (Bailin et al., 1999).

In contrast, the CT literature also shows a lack of consensus over a definition of CT due to the diversity of philosophical (e.g., Ennis, 2018; Facione, 1990; Paul & Nosich, 1993) and psychological (Halpern, 2003; Lai, 2011). approaches

and concepts. A widely cited conceptualization of CT is the one proposed by Ennis (2018), who defines it as "reflective and reasonable thinking focused on deciding what to believe or do, along with its expanded development of the dispositions and skills involved in such decisions". To create some consensus among specialists, a panel of experts from the American Psychological Association (APA, 1990, p. 3) proposed a definition of CT as the "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based". Many studies use this as a reference (APA, 1990; Facione, 1990).

As an alternative to this lack of conceptual consensus, some researchers choose to define CT by extension, that is, specifying its constitutive skills (Fisher, 2009). This approach is evident in the APA (1990) panel's definition, which mentions the skills of interpretation, analysis, evaluation, inference, judgment, and self-regulation. Ennis's (2018) expanded development also mentions decision-making. These examples or the extreme case of the so-called national plan for CT assessment (Paul & Nosich, 1993), which proposed a long list of 88 CT skills, evidence the lack of consensus on a definition of CT. However, some skills (e.g., analysis, reasoning, problem-solving, decision-making) and some dispositions (e.g., open-mindedness) can be considered predominant (Lai, 2011).

The different CT assessment instruments, by their functional and practical nature, tend to implicitly assume the extensive definition of CT, as each instrument usually specifies the skills it assesses. However, since the evaluation instruments are more specific than the definitions, here too, a lack of consensus is evident, highlighting the conceptual complexity of the CT construct. For example, the Cornell Critical Thinking Tests (CCTT) level X (Ennis et al., 2005a) assesses five dimensions (induction, deduction, observation, credibility, and assumptions), and the Critical Thinking Assessment test (HCTA) of Halpern (2007, 2010) assesses five skills (argument analysis, hypothesis-testing, probability and uncertainty, problem-solving, and verbal reasoning). In sum, the different skill terminologies, the unequal number of skills considered by each instrument (from 88 to 2), and the skills grouped into categories in some tools (those that offer a broader set of skills) are further examples of complexity, justifying the functionality of improving the organization of the CT field.

Some taxonomies of synthesis have been proposed to address this complexity and reduce the lack of consensus. For example, Dwyer et al. (2014) developed an integrated framework of educational objectives, cognitive processes (reflective judgment and self-regulation and meta-cognition functions), and CT skills (analysis, evaluation, and inference), including memory and comprehension as necessary processes to apply CT. Two recently developed taxonomies present a theoretical framework that organizes CT into four dimensions with significant coincidences between them. Manassero-Mas and Vázquez-Alonso (2019) proposed four basic dimensions of CT (creativity, reasoning and argumentation, complex processes, and evaluation and judgment), each containing multiple categories and subcategories (e.g., deductive, inductive, abductive, and statistical thinking; problem-solving and decision-making; assumptions, rules, dispositions). In a similar vein, Fisher (2021) also organized CT skills into four basic dimensions (interpretation, analysis, evaluation, and self-regulation), whose contents overlap broadly with the previous taxonomy.

In summary, the CT literature shows different conceptualizations among specialists, so to

avoid misconceptions, we used the taxonomy of Manassero-Mas and Vázquez-Alonso (2019) as a general reference. According to the authors, the term CT is fundamental and consists of four dimensions, each containing multiple specific thinking skills and other associated concepts (dispositions and rules of attitude). The taxonomy also reflects most of the CT skills involved in most CT assessment instruments. Finally, despite the discrepancies presented, all the authors agree on the educational importance of CT.

The evaluation of critical thinking

CT can be taught and learned, multiple CT teaching programs with varied orientations and practices have attempted to teach CT for decades (Follmann et al., 2018; Saiz, 2017; Swartz et al., 2013). In addition, recommendation 12 of the APA expert statement (Facione, 1998) endorsed complementing the teaching of CT with its frequent and explicit evaluation, both diagnostic and summative (Recommendation 13), and using valid, reliable and equitable instruments which currently are obvious features in the construction of tests (Muñiz & Fonseca-Pedrero, 2019).

The need to evaluate is justified by Ennis (2018) on the following grounds: diagnosing the students' CT skills level, providing feedback on progress, motivating learning of CT, informing teachers about teaching methods, investigating CT, counseling on the choice of studies, and stimulating educational institutions to report their results. The evaluation of CT is a necessity and significant support for improving its teaching, but it requires the construction of appropriate evaluation instruments to achieve valid and reliable results.

The specialized literature offers numerous tests to assess CT and, although most focus on a few CT skills (e.g., Facione et al., 1998; Halpern, 2010; Rivas & Saiz, 2012; Watson & Glaser, 2002), others are broader (Madison, 2004). The analysis of the skills included in the CT assessment tests provides an overview of the CT skills synthesized in the CT taxonomies mentioned (Ennis & Chattin, 2018; Fisher, 2021; Manassero-Mas & Vázquez-Alonso, 2019). However, CT teaching programs that have proven their effects through empirical evaluation studies are the exception rather than the rule (Saiz, 2017). Lipman's (1982) Philosophy for Children program has been repeatedly evaluated (Colom et al., 2014), while others, such as thought-based learning (Swartz et al., 2013), have only been evaluated occasionally, and others, such as the reasoning program (Walton & Macagno, 2015) still lack evaluation.

The vast majority of CT assessment instruments target at adults and university students, and there are hardly any specific tests for young students, although the Cornell tests (X, Y, Z) are partially adaptable to different ages (Ennis & Millman, 2005a, 2005b), and other proposals require a consolidated development (Lopes et al., 2018). In addition, the review of Aktoprak and Hursen (2022) shows a great lack of research on CT in primary education and a predominance of qualitative research methodologies in the few existing works. Therefore, Gelerstein et al., 2016; Lai, 2011; Meng, 2016; Pérez-Morán et al., 2021; Sierra et al., 2010 propose guiding studies towards quantitative research methodologies that complement qualitative research methods and strengthening the evaluation of CT with reliable measurement tests. Also, the differences between men and women have also been rarely investigated, although the study by Sierra et al. shows no significant differences.

In sum, the educational development of CT has been unequal among the different educational levels (frequent in university and rare in the lower educational levels) and within the contents (teach-

ing of CT has predominated, whereas reliable evaluation of CT is scarce, especially in younger students). These shortcomings justify this study's attention to the assessment of CT in young people, focused on specific skills appropriate and functional to their age range and contributing to drawing attention to the evaluation of CT in the early educational stages.

This study also builds on the development and evaluation of CT through the development of item banks on thinking skills for elementary school students (Manassero-Mas & Vázquez-Alonso, 2020a, 2020b). Based on the previous milestones and the application of psychometric recommendations to develop reliable tests (Fernández et al., 2010; Muñiz & Fonseca-Pedrero, 2019), a 48-item test that evaluates six thinking skills was validated. It is applied here to diagnose and highlight the thinking of primary school students. Its validity and reliability have been presented elsewhere (Manassero-Mas & Vázquez-Alonso, in press).

Consequently, the objectives of this study are: to quantitatively diagnose CT skills in 6th grade primary school children, present the normative data of the instrument, and compare the mastery of the skills in primary school children.

Method

Participants

The sample was comprised of 655 sixth-grade students (322 boys and 335 girls) with an average age of 11.16 years, who attended fourteen different schools in two Spanish communities (Catalonia, 42.6% and Balearic Islands, 57.4%), located in different towns (large, medium, small) of varied social contexts (upper, middle or lower class). Approximately half the participants studied in public schools (42.3%), and the other half (57.7%) in semi-private schools. All schools were selected for their favorable attitude towards critical thinking education. The students participated in this study in their own school groups, completing the thinking test as an assessment activity in the classroom under their teacher's direction.

Instrument

The test "Retos de Pensamiento" (RdP_EP6 [Thinking Challenges test]) applied in this study evaluates six CT skills: prediction and logical reasoning (reasoning dimension), comparison (creativity dimension), classification (evaluation dimension), and decision-making and problem-solving (complex processes dimension). These skills were agreed on with the schools participating in this study based on the skills adaptation to age and usual learning in sixth grade (EP6). The test items were designed using the criteria of readability, comprehensibility, balance on the cognitive demand of each item, the students' cognitive development, and the approach of facing a motivating and exciting challenge (Table 1).

Each test item was assigned to the skill most congruent with its content. For example, classification skill evaluates the ability to group or separate different elements according to their common or differential features. Prediction and comparison evaluate the ability to verify a logical conclusion through inductive reasoning or the creative contrast of several statements, respectively. Decision-making/problem-solving measures the ability to identify the best decisions/solutions in a particular situation, and logical reasoning evaluates simple (simple syllogism) and complex deductive ability (several pieces of information or conclusions are involved simultaneously).

The items propose a variety of scenarios and situations that communicate information by

various means of representation (verbal, numerical, and figurative). One or more questions are asked, whose cognitive demand is adjusted to the students' skill and age expected average, posing authentic and motivating thinking challenges (see a sample in the appendix). The content of the items is independent of the curricula of the school subjects (for example, they do not propose numerical calculations), so the correct answer does not require previous school knowledge but only applying elemental skills to the information presented. Therefore, the applied test, RdP_EP6, is cultural-free; that is, its challenges are not mediated by social, familiar and academic knowledge as many thinking tests are. For example, the Science CT test requires knowledge of the primary science curriculum to answer correctly (Mapeala & Siew, 2015).

The RdP_EP6 response formats are mostly closed (four items require a short open answer) because this allows for a standardized, fast, valid, and reliable evaluation of each thinking skill and for developing diagnostic baselines to compare research, programs, and teaching methodologies. The reliability values of the six skill scales and the total test (Table 1) correspond to the empirical factors obtained by procedures described below (unweighted least squares [ULS], Manassero-Mas & Vázquez-Alonso, in press).

Procedures

The RdP_EP6 was applied to the participants in their class group by their teachers as a regulated ordinary evaluation to stimulate the students' effort and motivation. The application followed standardized guidelines using digital devices with no time limit for the answers (usually completed in a class period).

Correct answers received one point, incorrect answers received zero points, and no corrections were applied to random answers. The score of each skill is the sum of the correct answers in the items that comprise it, and the overall score is the sum of all the correct answers (estimation of the students' overall CT).

The validity of the content of the RdP_EP6 is based on the credibility of the specialized publications consulted for the original items (Ennis &

Table 1
Specifications of the test applied (RdP_EP6) in this study to evaluate thinking skills in the sixth grade of Primary Education EP6.

Thinking skills	Source	Type	Items	Reliability (ORION*)
Prediction (PREDIC)	Ennis & Millman. 2005a	Verbal	9	.86
Comparison (COMP)	Ennis & Millman. 2005a	Verbal	7	.74
Classification (CLAS)	Author elaboration **	Figurative	6	.91
Problem-solving (PROB)	Halpern (2010)	Verbal	6	.81
	Author elaboration **	Figurative	4	
Decision-Making (DECIS)	Author elaboration **	Mixed	9	.86
Logical reasoning (LOG-RA)	Ennis & Millman. 2005b	Verbal	7	.86
Total			48	(Alpha) .85

* Overall Reliability of fully-Informative prior Oblique N-Expected a Posteriori
** Translated and adapted from open materials of https://www.criticalthinking.com

Millman, 2005a, 2005b; Halpern, 2010), the items prepared by the authors (https://www.pensamientocritico.com), and the researchers' professional judgment for the consensual selection of the items. The criteria for item selection were the best fit between the item's content and the represented skill and between the item's cognitive demand and the students' cognitive level.

Analysis of results

The data of individual scores were processed with SPSS (25). The validity and reliability of the test were presented extensively (Manassero-Mas & Vázquez-Alonso, in press). They were calculated with the program Factor 12.01.02 (Ferrando & Lorenzo-Seva, 2017, 2018; Lorenzo-Seva & Ferrando, 2019), which applies a robust method of unweighted least squares (ULS) based on tetrachoric correlations, appropriate for dichotomous test scores, exploratory factor analysis (EFA), confirmatory factor analysis (CFA) extract factors with ULS and Promin rotation and evaluate reliability using various indices, such as ORION and Cronbach's alpha (Table 1).

The evaluation of the differences among groups calculates the degree of significance of the differences among groups (ANOVA). The effect size statistic (ES, d) measures the magnitude of the differences in standardized units of deviation, independent of the sample size and the test applied, unlike the degree of statistical significance (Funder & Ozer, 2019; Schäffer & Schwarz, 2019).

The central issue of the ES is to determine whether or not an effect is relevant, for which conventional reference points are usually applied, which vary according to the field of study (Cohen, 1988; Rosenthal, 1996; Ventura-León, 2018). Educational research often reports ESs lower than other disciplines; for example, the meta-analysis of Hattie (2009) adopts $d > .40$ as a reference of the practical relevance of the educational effect and $d > .60$ is considered large. In this study, practical educational relevance was attributed to $d > .20$ because the probability is usually already statistically significant, and the following references were adopted for ES: Small ($d < .20$), medium (.20 - .30), moderate (.30 - .50), and large (> .50).

Results

The overall results of the 48 items that make up the RdP_EP6 are summarized in Table 2. The global average of the 48 items is .492, indicating that the test has a medium difficulty rate, very close to 50% of correct answers. In addition, there are six very difficult items (hit rate less than .30), and five very easy items (hit rate greater than .70), so 81% of the items have medium difficulty indexes included in the central range (.30 - .70).

Table 3 presents the descriptive results of the scores in the six thinking skills evaluated by the RdP_EP 6, obtained by adding the correct responses to the items that are part of each skill. As the number of items for each skill is different, the means obtained are not directly comparable. However, taking as a reference the central point of the scale of each skill, the results show that the prediction, classification, and problem-solving scales have means above their midpoint, whereas the comparison, decision-making, and logical reasoning scales obtain means below their midpoint. Hence, the former skills obtain overall hits above 50% (the easiest), whereas the latter ones obtain success rates below 50% (more difficult for the students).

Table 2
Descriptive statistics of the 48 items of the RdP_EP6 test (N = 655).

Variables		Mean	SD	Standard Error	95% Confidence interval of the mean	
					Lower limit	Upper limit
V1	PREDIC1	.62	.48	.02	.59	.66
V2	PREDIC2	.43	.50	.02	.39	.47
V3	PREDIC3	.50	.50	.02	.46	.54
V4	PREDIC4	.40	.49	.02	.36	.44
V5	PREDIC5	.79	.41	.02	.76	.82
V6	PREDIC6	.74	.44	.02	.70	.77
V7	PREDIC7	.71	.45	.02	.67	.74
V8	PREDIC8	.38	.49	.02	.34	.42
V9	PREDIC9	.64	.48	.02	.60	.67
V10	COMPA1	.44	.50	.02	.40	.48
V11	COMPA2	.56	.50	.02	.53	.60
V12	COMPA3	.43	.50	.02	.39	.47
V13	COMPA4	.52	.50	.02	.48	.56
V14	COMPA5	.50	.50	.02	.46	.54
V15	COMPA6	.50	.50	.02	.46	.54
V16	COMPA7	.36	.48	.02	.32	.39
V17	CLASIF1	.64	.48	.02	.60	.67
V18	CLASIF2	.55	.50	.02	.51	.59
V19	CLASIF3	.56	.50	.02	.52	.60
V20	CLASIF4	.65	.48	.02	.62	.69
V21	CLASIF5	.66	.47	.02	.63	.70
V22	CLASIF6	.64	.48	.02	.60	.68
V23	PROBL1	.65	.48	.02	.61	.69
V24	PROBL2	.61	.49	.02	.57	.65
V25	PROBL3	.57	.49	.02	.53	.61
V26	PROBL4	.32	.47	.02	.29	.36
V27	PROBL5	.73	.45	.02	.69	.76
V28	PROBL6	.73	.44	.02	.70	.77
V29	DECIS1	.35	.48	.02	.31	.39
V30	DECIS2	.28	.45	.02	.25	.32
V31	DECIS3	.23	.46	.02	.26	.33
V32	DECIS4	.33	.47	.02	.29	.36
V33	DECIS5	.43	.49	.02	.39	.46
V34	DECIS6	.19	.39	.02	.16	.22
V35	DECIS7	.15	.35	.02	.118	.17
V36	DECIS8	.65	.48	.02	.61	.68
V37	DECIS9	.46	.50	.02	.43	.50
V38	PROBL9	.21	.41	.02	.18	.24
V39	PROBL10	.43	.49	.02	.39	.46

Variables		Mean	SD	Standard Error	95% Confidence interval of the mean	
					Lower limit	Upper limit
V40	PROBL11	.54	.50	.02	.50	.57
V41	PROBL12	.38	.49	.02	.34	.41
V42	LOGIC1	.54	.50	.02	.50	.58
V43	LOGIC2	.55	.50	.02	.52	.59
V44	LOGIC3	.30	.47	.02	.27	.34
V45	LOGIC4	.59	.50	.02	.55	.63
V46	LOGIC5	.24	.42	.02	.20	.27
V47	LOGIC6	.57	.49	.02	.54	.61
V48	LOGIC7	.33	.47	.02	.29	.36

Table 3
Descriptive statistical results of the six thinking skills and the total score evaluated with the RdP_EP6 test.

Skills	Items	Mean	SD	Standard Error	95% Confidence interval of the mean		Minimum	Maximum
					Lower limit	Upper limit		
PREDICTION9	9	5.21	1.92	.07	5.06	5.35	0	9
COMPARISON7	7	3.31	1.42	.05	3.20	3.42	0	7
CLASSIFICATION6	6	3.70	1.89	.07	3.56	3.85	0	6
PROBLEMS10	10	5.17	2.14	.08	5.00	5.33	0	10
DECISIONS9	9	3.12	1.85	.07	2.98	3.27	0	9
LOGIC7	7	3.12	1.63	.06	3.00	3.25	0	7
TOTAL48	48	23.64	6.86	.27	23.11	24.16	10	45

The test's average total score (23.64) is close to 24, which marks the central point of the overall score, reflecting the intermediate global difficulty, close to 50%, in direct scores of the complete test. Also, the table indicates that the responses in all the skills reach the minimum (0) and maximum scores, which means that some students did not answer any item of the skill correctly, but also that some students answered all the items of each skill correctly. Regarding the global test, the minimum score achieved is 10 correct answers, and the maximum score is 45 correct answers, much closer to the possible maximum score (48) than the minimum score (10) regarding the possible minimum score (0 points).

Scale and normative data of the test

Table 4 presents the frequency distribution of the total RdP_EP6_48 scores obtained by the sample of students. The range extends from the minimum score of 10 correct answers to the maximum score of 45 correct answers. The mean score is 23.64 (Table 2), the median is 22, and the mode is 19.

The standardization of these scores in quartiles shows an asymmetric and distorted curve towards the highest scores because the highest quartile includes from score 28 to the maximum score of 45 (half the range of the scores obtained). This range is practically identical to the range of scores in the lower three quartiles (from the minimum score of 10 to score 27). Similarly, the distribution

Table 4
Standardized distribution of total RdP_EP6 test scores in the sample of primary school students.

Points	N	%	Percentiles	Quartiles
10	5	0.8		
11	9	2.1		
12	7	3.2		
13	9	4.6		
14	20	7.6	10	
15	28	11.9		
16	24	15.6		
17	29	20	20	
18	26	24		25
19	48	31.3	30	
20	37	36.9		
21	40	43.1	40	
22	37	48.7	50	50
23	33	53.7		
24	30	58.3	60	
25	23	61.8		
26	33	66.9		
27	25	70.7	70	
28	17	73.3		75
29	27	77.4		
30	27	81.5	80	
31	28	85.8		
32	25	89.6	90	
33	11	91.3		
34	12	93.1		
35	7	94.2		
36	12	96		
37	10	97.6		
38	3	98.0		
39	6	98.9		
40	3	99.4		
41	2	99.7		
44	1	99.8		
45	1	100		
Total	655			

of scores is strongly concentrated in the central percentile sections (between the 20th and 80th percentiles), which practically encompass one, two, or three different scores, whereas the lowest percentile section (10) comprises five different scores (between 10 and 14), and the highest percentile section (90) comprises 13 different scores (between 32 and 45).

The distribution of the scores on the six scales of the CT skills of the RdP_EP6 obtained by the sample of students is presented in Table 5. The range of the six scales is different, so the maximum scores vary according to the skill, from the shortest range of the classification skill (6) to the longest range of the problem-solving skill (10).

Gender differences in thinking skills

To evaluate the gender differences in thinking skills, we compared the scores obtained by the groups of boys and girls in all the variables of the RdP_EP6 considered in this study, which meet the conditions of normality, equality of variances, and sample similarity. The relevance of the differences between the two groups was measured with two statistics: the degree of significance of the differences (through ANOVA) and the ES of the differences (through Cohen's formula, as the two groups are similar in size). The ES was computed subtracting the girls' average to the boys'mean. Thus, positive differences indicate the boys' higher score, and negative differences indicate the girls' higher score.

Table 6 presents the results of the means and standard deviations for each of the 48 items that make up the test of the two compared groups of boys and girls, the two statistics assessing the differences, the degree of significance of the differences (p) and the ES of the differences (d), ordered from highest to lowest according to the ES.

Table 5
Distribution of scores on the six CT skill scales of the RdP_EP6 test.

Points	PREDIC9		COMPA7		CLASIF6		PROBL10		DECIS9		LOGIC7	
	N	%	N	%	N	%	N	%	N	%	N	%
0	5	0.8	7	1.1	29	4.4	4	0.6	31	4.7	33	5
1	18	3.5	54	9.3	72	15.4	14	2.7	102	20.3	78	16.9
2	43	10.1	142	31	103	31.1	49	10.2	143	42.1	140	38.3
3	66	20.2	167	56.5	94	45.5	91	24.1	131	62.1	126	57.6
4	79	32.2	140	77.9	88	58.9	109	40.8	93	76.3	135	78.2
5	134	52.7	102	93.4	101	74.4	118	58.8	80	88.5	97	93
6	135	73.3	40	99.5	168	100	74	70.1	44	95.3	37	98.6
7	106	89.5	3	100			88	83.5	21	98.5	9	100
8	51	97.3					70	94.2	7	99.5		
9	18	100					27	98.3	3	100		
10							11	100				
Total	655		655		655		655		655		655	

Table 6
Descriptive statistics of the 48 items of the RdP_EP6 test for boys and girls, the degree of significance and the effect size of the group differences ordered by effect size.

Variables		Boys		Girls		P-Sig.	Effect size
		Mean	SD	Mean	SD		
V38	PROBL9	.25	.44	.17	.38	.009	.195
V35	DECIS7	.17	.37	.13	.33	.144	.114
V13	COMPA4	.55	.50	.50	.50	.189	.100
V31	DECIS3	.31	.46	.28	.45	.419	.066
V4	PREDIC4	.41	.49	.39	.49	.750	.041
V12	COMPA3	.44	.50	.42	.49	.725	.040
V46	LOGIC5	.24	.43	.23	.42	.611	.024
V44	LOGIC3	.31	.46	.30	.46	.827	.022
V7	PREDIC7	.71	.45	.71	.46	.957	.000
V3	PREDIC3	.49	.50	.50	.50	.784	-.020
V39	PROBL10	.42	.49	.43	.50	.837	-.020
V23	PROBL1	.64	.48	.65	.48	.789	-.021
V26	PROBL4	.32	.47	.33	.47	.731	-.021
V27	PROBL5	.72	.45	.73	.44	.786	-.022
V2	PREDIC2	.42	.49	.44	.50	.452	-.040
V15	COMPA6	.49	.50	.51	.50	.726	-.040
V41	PROBL12	.37	.48	.39	.49	.554	-.041
V5	PREDIC5	.78	.42	.80	.40	.435	-.049
V48	LOGIC7	.31	.46	.34	.47	.449	-.065
V28	PROBL6	.72	.45	.75	.43	.290	-.068
V42	LOGIC1	.52	.50	.56	.50	.277	-.080

Variables		Boys		Girls		P-Sig.	Effect size
		Mean	SD	Mean	SD		
V10	COMPA1	.42	.49	.46	.50	.258	-.081
V1	PREDIC1	.60	.49	.64	.48	.238	-.082
V22	CLASIF6	.62	.49	.66	.47	.210	-.083
V30	DECIS2	.26	.44	.30	.46	.269	-.089
V43	LOGIC2	.53	.50	.58	.49	.142	-.101
V17	CLASIF1	.61	.49	.66	.47	.182	-.104
V36	DECIS8	.62	.49	.67	.47	.158	-.104
V6	PREDIC6	.71	.45	.76	.43	.185	-.114
V14	COMPA5	.47	.50	.53	.50	.172	-.120
V25	PROBL3	.54	.50	.60	.49	.091	-.121
V8	PREDIC8	.35	.48	.41	.49	.121	-.124
V29	DECIS1	.32	.47	.38	.49	.125	-.125
V21	CLASIF5	.63	.48	.69	.46	.136	-.128
V34	DECIS6	.16	.37	.21	.41	.152	-.128
V40	PROBL11	.50	.50	.57	.50	.101	-.140
V33	DECIS5	.39	.49	.46	.50	.104	-.141
V45	LOGIC4	.55	.50	.62	.49	.055	-.141
V16	COMPA7	.32	.47	.39	.49	.077	-.146
V20	CLASIF4	.62	.49	.69	.46	.047	-.147
V19	CLASIF3	.52	.50	.60	.49	.030	-.162
V9	PREDIC9	.59	.49	.68	.47	.013	-.188
V18	CLASIF2	.50	.50	.60	.49	.013	-.202
V24	PROBL2	.56	.50	.66	.48	.013	-.204
V32	DECIS4	.28	.45	.38	.49	.008	-.213
V37	DECIS9	.41	.49	.52	.50	.004	-.222
V47	LOGIC6	.52	.50	.63	.48	.005	-.224
V11	COMPA2	.50	.50	.62	.49	.002	-.242

The main finding comparing boys and girls is that most of the differences obtained in all the 48 items show that girls score higher than boys in 39 items (negative ES), and boys score higher in only eight of the remaining items (positive ES). This result indicates that girls in the sixth grade of primary education have, on average, better CT skills than boys.

The second finding in Table 6 is that the differences between boys are low and mostly non-significant. Indeed, all the differences between boys and girls calculated through the ES are less than .30, and among the highest that obtain negative values favoring the girls, only six items exceed the value .20. Similarly, only nine items reach a significance level of $p < .05$ (of which four items reach $p < .01$). In sum, the significance and ES of the differences between boys and girls are small.

The results obtained for the gender differences between primary school boys and girls in the six skill variables and the total score of the questionnaires confirm and reinforce the patterns and trends found for the 48 items of the test, given the additive nature of the skill scales (Table 7).

All the differences in the six skills and the total score favor the girls (negative), which shows the overwhelming dominance of girls in almost all the items. Although most scores achieve statistically significant differences, the ES of the differences remains low. The largest gender difference is in the total score, and the smallest is in problem-solving skills.

Table 7
Descriptive statistics of the six CT skills and the total score of the RdP_EP6 test (means and standard deviations) in the group of boys and girls, with the degree of significance and the effect size of the group differences.

Skills	Boys		Girls		P-Sig.	Effect size
	Mean	SD	Mean	SD		
PREDICTION9	5.06	1.91	5.35	1.93	.051	-.151
COMPARISON7	3.19	1.41	3.43	1.42	.035	-.170
CLASSIFICATION6	3.49	1.89	3.90	1.87	.006	-.218
PROBLEMS10	5.04	2.18	5.28	2.09	.151	-.112
DECISIONS9	2.92	1.80	3.32	1.88	.006	-.217
LOGIC7	2.98	1.67	3.26	1.58	.025	-.172
TOTAL48	22.69	6.89	24.54	6.72	.001	-.272

Discussion and conclusions

The main objective of this study is to diagnose the level of CT skills in a large sample of sixth-grade students of Primary Education (11 years) through the RdP_EP6 test, which evaluates six CT skills (prediction, comparison, classification, problem-solving, decision-making, and logical reasoning). The results indicate that the students reach an intermediate level of mastery of CT skills (about 50% of correct answers in the global test) on the cognitive demands of the test items, a first reference of mastery for this sample and this test. Concerning the relative mastery of the different skills assessed in the test, the students show a greater relative mastery of prediction, classification, and problem-solving skills (scores above the midpoint of each measurement scale), whereas comparison, decision-making, and logical reasoning skills have relatively lower scores (below the midpoint of each measurement scale). These results are complemented with the psychometric evaluation of the test and the scales of the six skills, which can serve as a global reference framework for the expansion of the test's standardization with different samples from other contexts and places, contributing to the development of a valid and reliable test (Manassero-Mas & Vázquez-Alonso, in press).

CT studies in primary education are few and qualitative (Gelerstein et al., 2016; Lai, 2011; Meng, 2016; Pérez-Morán et al., 2021). In addition, there is a lack of specific tests to evaluate CT in youngsters and there are even fewer studies evaluating skills which do not assess students' real mastery of CT skills. For example, Lopes et al. (2018) developed a qualitative test for students from 12 to 19 years old, and Pérez-Moran et al. (2021) did so quantitatively, but they did not value the real mastery of the students' performance. In short, there is a lack of quantitative studies that can serve as a reference to assess the domain of CT reflected in the scores of the skills evaluated in primary education. This prevents contrasting the scope and value of the results obtained in the sample of this study with other equivalent samples evaluated with different instruments. Thus, these results are pioneer in serving as a precedent

and diagnostic reference for subsequent studies and they contribute to filling the gaps, although the test's valuation is pending future confirmation.

The most notable finding is the girls' higher level of CT skills in most test items, the six skills and the total CT score. The differences in favor of girls are statistically significant in comparison, classification, decision-making, logical reasoning and, of course, the total CT score. In sum, although the magnitudes of the differences are not large, the statistically significant superiority of girls over boys in CT skills constitute a consistent and solid trend. This supports girls' better performance instead of ratifying a hypothesis of similarity of the two groups in primary school students (Sierra et al., 2010) or the differences in older students, obtained with statistics inappropriate to the group size (Lopes et al., 2018).

Girls' better CT mastery suggests two interesting facts. The first refers to the justification and explanation of this differential result because if boys and girls have mostly attended the same school together, in the same classes, and with the same teachers, there is no evidence to attribute the differences to cultural or educational variables. Thus, the explanatory parameters could be within the framework of the evolutionary differences between boys and girls.

An additional interesting issue is related to the hypothesis of similarity of men and women presented in the literature of differential psychology, where it is still considered that spatial mental rotation is the only capacity that presents large empirical differences favoring men, controlling for the educational and cultural background (Jäncke et al., 2018). Item V2638-PROBL9 of the RdP_EP6 makes a cognitive demand that involves imagining the rotation of a cube to give the correct answer, and its result of gender differences (Table 6) presents the greatest magnitude of the differences favoring boys (d = 0.195), consistent with differential psychology's prediction about spatial rotation. This result confirms differential psychology research on spatial rotation and further supports the validity and reliability of the RdP_EP6.

References

Aktoprak, A., & Hursen, C. (2022). A bibliometric and content analysis of critical thinking in primary education. *Thinking Skills and Creativity, 44*, 101029. https://doi.org/10.1016/J.TSC.2022.101029

Almerich, G., Suárez-Rodríguez, J., Díaz-García, I., & Orellana, N. (2020). Estructura de las competencias del siglo XXI en alumnado del ámbito educativo. Factores personales influyentes [Structure of 21st century competences in students the sphere of education. Influential personal factors]. *Educación XX1, 23*(1), 45-74. https://doi.org/10.5944/educXX1.23853

American Psychological Association APA (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report)*. https://philarchive.org/archive/faccta

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies, 31*(3), 285-302. https://doi.org/10.1080/002202799183133

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Academic Press.

Colom, R., García Moriyón, F., Magro, C. & Morilla, E. (2014). The long-term impact of philosophy for children: A longitudinal study (Preliminary results). *Analytic Teaching and Philosophical Praxis, 35*(1), 50-56. https://journal.viterbo.edu/index.php/atpp/article/view/1129

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity, 12*, 43-52. https://doi.org/10.1016/J.TSC.2013.12.004

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi, 37*, 165-184. https://doi.org/10.1007/

s11245-016-9401-4

Ennis, R. H. & Chattin, G. S. (2018). An annotated list of critical thinking tests. http://criticalthinking.net/wp-content/uploads/2018/01/An-Annotated-List-of-English-Language-Critical-Thinking-Tests.pdf

Ennis, R. H., & Millman, J. (2005a). *Cornell Critical Thinking Test Level X*. The Critical Thinking Company.

Ennis, R. H., & Millman, J. (2005b). *Cornell Critical Thinking Test Level Z*. The Critical Thinking Company.

European Union (2014). *Key competence development in school education in Europe. KeyCoNet's review of the literature: A summary*. European Schoolnet. http://keyconet.eun.org

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations*. American Philosophical Association. https://eric.ed.gov/?id=ED315423

Facione, P. A. (1998). *Insight assessment*. www.insightassessment.com

Facione, P. A., Facione, N. C., Blohm, S.W., Howard, K., & Giancarlo, C. A. F. (1998). *California Critical Thinking Skills Test: Manual (Revised)*. California Academic Press.

Fernández, A., Pérez, E., Alderete, A. M., Richaud, M. C., & Fernández Liporace, M. (2010). ¿Construir o adaptar tests psicológicos? Diferentes respuestas a una cuestión controvertida. *Revista Evaluar, 10*(1). https://doi.org/10.35670/1667-4545.v10.n1.459

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*, 236-240. https://doi.org/10.7334/psicothema2016.304

Ferrando, P. J., & Lorenzo-Seva U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*, 762-780. https://doi.org/10.1177/0013164417719308

Fisher, A. (2009). Critical thinking. An introduction. Cambridge University Press.

Fisher, A. (2021). On what critical thinking is. In J. A. Blair (Ed.), *Studies in critical thinking* (2nd ed., pp. 7-26). University of Windsor. https://doi.org/10.22329/wsia.08.2019

Follmann, D., Mattos, K. R. C., & Güllich, R. I. da C. (2018). Teaching strategies of sciences and the promotion of critical thinking in Portugal. *Tecné, Episteme y Didaxis* (Extraordinario, Octavo Congreso Internacional de formación de Profesores de Ciencias para la Construcción de Sociedades Sustentables). https://revistas.pedagogica.edu.co/index.php/ESD/article/view/8789

Fullan, M., & Scott, G. (2014). *Education PLUS*. Collaborative Impact SCT. https://michaelfullan.ca

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Gelerstein, D., del Río, R., Nussbaum, M., Chiuminatto, P., & López, X. (2016). Designing and implementing a test for measuring critical thinking in primary school. *Thinking Skills and Creativity, 20*, 40-49. https://doi.org/10.1016/J.TSC.2016.02.002

Halpern , D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. Laurence Erlbaum Associates.

Halpern, D. F. (2007). *Halpern Critical Thinking Assessment using everyday situations: Background and scoring standards*. Claremont McKenna College.

Halpern , D. F. (2010). *Manual Halpern Critical Thinking Assessment*. Schuhfried GmbH.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

International Society for Technology Education. (2003). *National educational technology standards for teachers: Preparing teachers to use technology*. ISTE.

Jäncke, L. (2018). Sex/gender differences in cognition, neurophysiology, and neuroanatomy. *F1000Research, 7*,805. https://doi.org/10.12688/

f1000research.13917.1

Krathwohl, D. (2002). A revision of Bloom's taxonomy: An Overview. *Theory into Practice, 41*, 212-218. https://doi.org/10.1207/s15430421tip4104_2

Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's research reports, 6*, 1-49.

Lipman, M. (1982). Philosophy for Children. *Thinking: The Journal of Philosophy for Children, 3*(3/4), 35-44. https://doi.org/10.5840/thinking1982339

Lopes, J., Silva, H., & Morais, E. (2018). Teste de pensamento crítico para estudantes dos ensinos básico e secundário [Critical thinking test for elementarý and secondary students]. *Revista de Estudios e Investigación en Psicología y Educación, 5*(2), 82-91. https://doi.org/10.17979/reipe.2018.5.2.3339

Lorenzo-Seva, U., & Ferrando, P. J. (2019). Robust promin: A method for diagonally weighted factor rotation. *LIBERABIT, Revista Peruana de Psicología, 25*, 99-106. https://doi.org/10.24265/liberabit.2019.v25n1.08

Madison, J. (2004). *James Madison critical thinking course.* The Critical Thinking Co.

Manassero-Mas, M. A., & Vázquez-Alonso, A. (2019). Taxonomía de las destrezas de pensamiento: una herramienta clave para la alfabetización científica [Taxonomy of thinking skills: A key tool for scientific literacy]. In M. D. Maciel & E. Albrecht (Org.), *Ciência, Tecnologia & Sociedade: Ensino, Pesquisa e Formação* (pp. 17-38). UNICSUL.

Manassero-Mas, M. A., & Vázquez-Alonso, A. (2020a). Evaluación de destrezas de pensamiento crítico: Validación de instrumentos libres de cultura [Assessment of critical thinking skills: Validation of free-culture tools]. *Tecné, Epistemé y Didaxis, 47*, 15-32. https://doi.org/10.17227/ted.num47-9801

Manassero-Mas, M. A., & Vázquez-Alonso, A. (2020b). Las destrezas de pensamiento y las calificaciones escolares en educación secundaria: Validación de un instrumento de evaluación libre de cultura [Thinking skills and school grades in secondary education: Validation of a culture-free assessment instrument]. *Tecné, Epistemé y Didaxis, 48*, 33-54. https://doi.org/10.17227/ted.num48-12375

Manassero-Mas, M. A., & Vázquez-Alonso, A. (in press). Assessment of critical thinking skills in primary education: Validation of Challenges of Thinking Test. *Thinking Skills and Creativity*.

Mapeala, R., & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth graders. *SpringerPlus, 4*(1). https://doi.org/10.1186/s40064-015-1535-0

Meng, K.H. (2016). Infusion of critical thinking across the English language curriculum: A multiple case study of primary school in-service expert teachers in Singapore. Ph.D. Thesis, University of Western Australia, Perth, Australia.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

National Education Association. (2012). *Preparing 21st-century students for a global society: An educator's guide to the "four Cs"*. National Education Association.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398

Organization for Economic Co-operation and Development. (2018). The future of education and skills. *Education 2030*. OECD Publishing. https://www.oecd.org/education/2030-project/contact

Paul, R., & Nosich, G. M. (1993). A model for the national assessment of higher order thinking. In R. Paul (Ed.), *Critical thinking: What every student needs to survive in a rapidly changing world* (pp. 78-123). Foundation for Critical Thinking.

Pérez-Morán, G., Bazalar-Palacios, J., & Arhuis-Inca, W. (2021). Diagnóstico del pensamiento crítico de estudiantes de educación primaria de Chimbote, Perú [Diagnosis of critical thinking of elementary school students in Chimbote, Peru]. *Revista*

*Electrónica Educare, 25(*1), 289-299. https://dx.doi.org/10.15359/ree.25-1.15

Piaget, J., & Inhelder, B. (1997). *Psicología del niño* [The psychology of the child]. Morata.

Rivas, S. F., & Saiz, C. (2012). Validación y propiedades psicométricas de la prueba de pensamiento crítico PENCRISAL. *Revista Electrónica de Metodología Aplicada, 17*(1), 18-34. https://reunido.uniovi.es/index.php/Rema/index

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37-59. https://doi.org/10.1300/J079v21n04_02

Saiz-Sánchez, C. (2017). *Pensamiento crítico y cambio* . Pirámide.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.00813

Shayer, M., & Adey, P. S. (Eds.) (2002). *Learning intelligence: Cognitive acceleration across the curriculum from 5 to 15 years*. Open University Press.

Sierra-Paz, J., Carpintero-Molina, E., & Pérez-Sánchez, L. (2010). Pensamiento crítico y capacidad intelectual. *Faísca Revista de divulgación científica sobre las altas capacidades intelectuales, 15*(17), 98-110. https://www.revistafaisca.es

Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan R., & Kallick, B. (2013). *El aprendizaje basado en el pensamiento. Cómo desarrollar en los alumnos las competencias del siglo XXI* [Thinking-based learning. Promoting quality student achievement in the 21st Century]. Ediciones SM.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). AHELO. *Assessment of higher education learning outcomes. Volume 1 - Design and implementation (Feasibility Study Report)*. OECD. http://www.oecd.org/education/skills-beyond-school

UNESCO. (2016). *Education 2030: Incheon declaration and framework for action for the implementation of sustainable development goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all*. https://unesdoc.unesco.org

Valenzuela, J. (2008). Habilidades de pensamiento y aprendizaje profundo. *Revista Iberoamericana de Educación, 46*(7), 1-9. https://doi.org/10.35362/rie4671914

Ventura-León, J. (2018). Otras formas de entender la d de Cohen [Other ways of understanding Cohen's d]. *Revista Evaluar, 18*(3). https://doi.org/10.35670/1667-4545.v18.n3.22305

Vincent-Lancrin, S., González-Sancho, C., Bouckaert, M., de Luca, F., Fernández-Barrerra, M., Jacotin, G., Urgel, J., & Vidal, Q. (2019). *Fostering students' creativity and critical thinking*. OECD. https://doi.org/10.1787/62212c37-en

Walton, D., & Macagno, F. (2015). A classification system for argumentation schemes. *Argument and Computation, 6*(3), 214-249. https://www.tandfonline.com/loi/tarc20

Watson, G., & Glaser, E. M. (2002). *Watson-Glaser Critical Thinking Appraisal-II Form E*. Pearson.

World Economic Forum (2021). *These are the top 10 job skills of tomorrow - and how long it takes to learn them*.