2024, Vol. 24, 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

Assessment of Critical Thinking Skills in Primary Education: Validation of Challenges of Thinking Test

Evaluación de destrezas de pensamiento crítico en Educación Primaria: Validación de la prueba Desafíos del Pensamiento

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

1 - University of the Balearic Islands, Palma, Spain.
2 - University of the Balearic Islands, Palma, Spain.

Recibido: 22/03/2024 **Revisado:** 05/06/2024 **Aceptado:** 08/06/2024

Abstract

The global demand for 21st century competencies raises critical thinking (CT) as a priority educational objective, which in turn projects the need for CT evaluation. The lack of CT assessment instruments for youngsters justifies the aim of this study: to develop a CT test for Primary Education and unveil its psychometric properties. The methodology follows the test development prescriptions through the elaboration of the six-skill CT test, the test application to primary sixth-graders and the confirmatory factor analysis on the answers. Starting from a 48-item test form, the empirical analysis confirms a six-factor structure, interpret the six empirical factors in face of the postulated CT skills, confirm a one-dimension structure for the whole test and for each of the six factors (except Comparison), and describe the goodness-of-fit psychometric parameters that support the reliability and validity of the 31-item test form. Finally, the properties, utility, limitations, and prospective improvements, developments and applications of the test for education and CT research are discussed.

Keywords: *student evaluation, critical thinking test, validity, reliability, primary education*

Resumen

La demanda global de competencias del siglo XXI plantea el pensamiento crítico (PC) como un objetivo educativo prioritario, lo que a su vez proyecta la necesidad de la evaluación del PC. La falta de instrumentos de evaluación del PC para jóvenes justifica el objetivo de este estudio: desarrollar un test de PC para Educación Primaria y presentar sus propiedades psicométricas. La metodología sigue las prescripciones de desarrollo de pruebas a través de la elaboración de una prueba de PC con seis destrezas, la aplicación de la prueba a estudiantes de sexto grado de primaria y el análisis factorial confirmatorio de las respuestas. Partiendo de un formulario de prueba de 48 ítems, el análisis empírico confirma una estructura de seis factores, interpreta los seis factores empíricos frente a las destrezas de PC postuladas, confirma una estructura unidimensional para toda la prueba y para cada uno de los seis factores (excepto Comparación), y describe los parámetros psicométricos de bondad de ajuste que respaldan la confiabilidad y validez de una forma de la prueba con 31 ítems. Finalmente, se discuten las propiedades, utilidad, limitaciones y posibles mejoras, desarrollos y aplicaciones de la prueba para la educación y la investigación en PC

Palabras clave: *evaluación de estudiantes, prueba de pensamiento crítico, validez, confiabilidad, educación primaria*

***Correspondence to:** Ángel Vázquez-Alonso. E-mail: angel.vazquez@uib.es. Address: Universidad de las Islas Baleares, Edificio Guillem Cifre de Colonya. Carretera de Valldemossa, km. 7.5, 07122 - Palma, Spain.
Tel. +34 971 17 30 75. FAX: 34 971 17 31 90.

How to cite: Manassero-Mas, A., M., & Vázquez-Alonso, Á.,(2024): Assessment of Critical Thinking Skills in Primary Education: Validation of Challenges of Thinking Test. *Revista Evaluar, 24*(2), 53-75. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar
Acknowledgments: EDU2015-64642-R (AEI/FEDER, UE) funded by the Spanish State Research Agency and the European Regional Development Fund.
Declaration of conflict of interest: none.
Participaron en la edición de este artículo: Stefano Macri, Eugenia Barrionuevo, Abigail Pérez, Florencia Ruiz, Jorge Bruera.

Introduction

Critical thinking is currently an overarching concept in psychology, philosophy, education and job. Since centuries philosophers adopted CT as their working tool to quality thinking that brings along high standards (precision, solidity, coherence, etc.) and to avoid error, fallacy and bias (ego-centrism and socio-centrism) (Ennis, 2018; Bailin et al., 1999; Facione, 1990). Psychologists developed CT as a set of higher order cognitive skills (e.g., inference, analysis, problem-solving, interpretation, creativity, decision-making, evaluation, etc.) along with a set of attitudinal dispositions (e.g., truth-seeking, self-confidence, curiosity, open-mindedness, etc.) that drive the adequate application of skills (Fisher, 2021; Halpern, 2003; Manassero-Mas et al., 2022).

Worldwide educational institutions and experts are claiming CT as a core part of the 21st century skills that citizens need to face the current challenges (globalization, accelerated development, ecological emergency, etc.) and their consequent personal, labor, and social impacts (Almerich et al., 2020; European Union, 2014; Fullan & Scott, 2014; International Society for Technology Education, n.d.; National Research Council, 2012; Organisation for Economic Co-operation and Development [OECD], 2018; UNESCO, 2016; Vincent-Lancrin et al., 2019). From the employers' view, most surveys reiterate CT at the top of skills required for future jobs (Whiting, 2021) and a key factor for people's success in the information age (Tremblay, 2013).

The mastery of CT is a key educational factor for significant and deep learning that characterizes educational excellence (Hattie, 2012; Valenzuela, 2008). In fact, CT aligns with Piaget's pioneering studies (Piaget & Inhelder, 1997) and the cognitive acceleration programs (Shayer & Adey, 2002), which have empirically demonstrated their significant impact on learning (effect size = 1.28), according to Hattie's (2009, 2012) meta-analysis of visible learning, which additionally assigns a large impact to some CT skills (meta-cognitive strategies, creativity, problem-solving, etc.).

The beneficial consequences of mastering CT for learning, personal development and success in job and social contexts justify the research attention to CT. However, teaching, assessing and making CT visible is still difficult for most elementary schools due to the lack of CT resources and tests for primary education, as justified below. In order to fill in these gaps, this study aims to develop a quantitative, valid and reliable CT test for primary education from an educational, diagnostic and evaluation perspective.

The conceptual framework on critical thinking

The literature research on CT developed along three basic lines since years: conceptualization, teaching, and evaluation, though each line has achieved a misbalanced development (Saiz, 2017).

The conceptualization of CT displays a lack of consensus, as the researchers display a big diversity of concepts and terms among that impede achieving a shared definition for CT (e.g., Ennis, 2018; Facione, 1990; Halpern, 2003; Paul & Nosich, 1993). The Ennis' (2018) conceptualization of CT, as reflective and reasonable thinking focused on deciding what to believe or do, and its expanded development in dispositions and skills involved in such decisions are widely cited. In the pursuit of consensus, a panel of experts from the American Philosophical Association (APA) proposed a definition of CT as the deliberate and self-regulating judgment for a specific purpose, employing evidence-based interpretation, analysis, evaluation, and inference, concepts, methods, criteria, and contexts to establish such judgments (Facione, 1990).

As an alternative way, some researchers choose to define CT by specifying its constitutive skills, yet this perspective has not achieved consensus either (Fisher, 2009). For instance, the aforementioned APA panel definition proposed the following skills: interpretation, analysis, evaluation, inference, interpretation, judgment, and self-regulation. On the other hand, the national plan for the assessment of CT proposed an 88-skill list, yet they were grouped in dimensions (Paul & Nosich, 1993).

Further, the practical function of CT assessment tests requires clearly specifying the skills they evaluate, so that the analysis of tests' specific skills might shed more light on CT conceptualization than CT definitions. However, the comparison of the different CT assessment tests shows the sets of skills are different across tests, again lacking relevant coincidences with each other. Thus, CT tests also lead to conceptual discrepancy and complexity about the CT construct along its constitutive skills. The different skills labels, the unequal number of test skills (ranging between 88 and 2), and the categories of skills add to the CT complexity (Manassero-Mas & Vázquez-Alonso, 2019).

Some synthetical taxonomies have been proposed to alleviate the lack of consensus on conceptualizing CT. Dwyer et al. (2014) developed an integrated framework of educational objectives, cognitive processes (reflective judgment and self-regulation, and meta-cognition) and CT skills (analysis, evaluation, and inference), with memory and comprehension as necessary processes for CT application. Two recent theoretical frameworks coincidentally organize CT into four similar dimensions. Manassero-Mas and Vázquez-Alonso (2019) proposed CT as the foundational construct, which develops along four dimensions (creativity, reasoning and argumentation, complex processes, and evaluation and judgment), each containing categories of thinking skills (e.g., deductive, inductive, abductive, statistical thinking, problem-solving, decision-making) and other associated concepts (assumptions, standards, attitudinal dispositions, norms). Similarly, Fisher (2021) organized CT skills across four basic groups (interpretation, analysis, evaluation, and self-regulation), whose contents widely overlap with the previous taxonomy.

All in all, the CT conceptualization shows important differences across authors, both theoretical and practical. The synthetical and integrative taxonomies provide a balanced view of the CT, echo the CT skills involved in most CT tests and avoid a dysfunctional alphabet soup in the field. Herein, the Manassero-Masa and Vázquez-Alonso' (2019) taxonomy will be used as an overall reference to frame the researched CT test.

Teaching and evaluating critical thinking

In spite of the conceptual disagreements, all CT experts endorse the assumption that thinking can be taught and learnt, which lead to the development of a variety of teaching and assessment programs (Follmann et al., 2018; Saiz, 2017; Swartz et al., 2013). The recommendations (12 and 13) of APA experts' statement (Facione, 1990) advocated frequent, explicit, diagnostic and summative evaluations of CT, through valid, reliable and equitable tools, currently obvious features of tests (Muñiz & Fonseca-Pedrero, 2019). Further, Ennis (2018) provided many reasons to assess CT: diagnostic of students' skills, feedback on program progress, motivation to learn CT, information about teaching, investigate CT, counsel about study choice and stimulation to report results.

The evaluation of CT requires the construction of appropriate tests to assess valid and reliable measurements, plus some of those tests have been developed. Most tests focus on assessing a few CT skills (e.g., Facione et al., 1998; Halpern, 2010; Rivas & Saiz, 2012; Watson & Glaser, 2002), yet some are broader and ambitious (Madison, 2004). The analysis of the CT skills included in the tests lead to synthesizing the mentioned taxonomies (Manassero-Mas & Vázquez-Alonso, 2019; Ennis, 2009; Fisher, 2021).

However, the programs that have proven their effects through empirical evaluation are the exception rather than the rule (Saiz, 2017). Lipman's (1982) philosophy for children has been repeatedly evaluated (Colom et al., 2014), while others have only been occasionally appraised (e.g., thinking-based learning, Swartz et al., 2013), and still others lack evaluations (e.g., the reasoning program, Walton & Macagno, 2015). So, this paper tries to construct a valid and reliable test that can be deemed functional for feedback on the educational programs of primary education.

The vast majority of CT assessment instruments address adults and university students, and there are hardly any tests for young students, though some items of the Cornell tests may be adaptable to young people (Ennis & Millman, 2005a, 2005b), and other proposals require further development (Lopes et al., 2018). The review of Aktoprak and Hursen (2022) diagnoses the scarcity and weaknesses of CT research in primary education, proposes increasing it by intensifying the evaluation of CT skills in educational projects, and points out to use reliable tests to complement the predominant qualitative methods in primary (e.g., Gelerstein et al., 2016). Thus, the new test developed here adheres to this proposal on quantitative and functional CT assessment for primary education.

In sum, the scant attention paid to the youngest students' teaching and assessment of CT lead to the inadequate of the previously mentioned tests for children. The growing importance of CT in education, as a key constituent of 21st century skills, points out a return to this situation and justifies the development of a new test to evaluate young people's CT. This aim involves the test content being adapted to the developmental ability of the primary students: focus on some specific, appropriate, and functional skills, adequate the item cognitive demand and make the test independent of the curricular knowledge. Further, the test must meet a balanced development of each of the dimensions of the CT taxonomy that has been adopted as a reference (Manassero-Mas & Vázquez-Alonso, 2019), to make thinking visible at early educational stages. Consequently, the objective of this study is to develop an assessment instrument to quantitatively diagnose the CT skills of young primary school students, investigate their relationship with learning (represented by school grades as empirically validated external criteria), and apply confirmatory analytical methods to support the psychometric validity and reliability of the instrument.

This study continues a work that has already developed some previous stages: the construction of a wide bank of Spanish items on CT skills, which allow the development of some pilot applications involving many items and small samples of sixth-graders, in order to classify the tested items according to their difficulty, their fit within CT skills and their mutual correlations (Manassero-Mas & Vázquez-Alonso, 2020a, 2020b). On this basis, a previous study developed and evaluated a test form that achieved hopeful results with sixth graders, yet the five-skill final test still left room for its validation improvement (Manassero-Mas & Vázquez-Alonso, 2024). Thus, applying the usual recommendations for test development to the former five-skill final test, a new 48-item and six-skill test was raised, as the starting point of this validation study (Ferrando et al., 2022; Muñiz & Fonseca-Pedrero, 2019). Likewise, the results of this 48-item test were applied as a diagnostic evaluation of the Spanish primary school students' thinking skills (Manassero-Mas & Vázquez-Alonso, 2023).

Methodology

A new 48-item and six-skill test (mentioned above) is the starting form of this new validation study through its application to assess the CT skills of elementary students. The methodology and results are presented here.

Participants

The complete 48-item test was applied to a varied convenience sample comprising 655 students (320 male and 335 female) sixth graders of Primary Education, aged between 10 and 13 years (average 11.1 years). The students attended fourteen public and public-funded schools, which were located in a variety of places in three regions. The schools were willing towards teaching and learning thinking and students were tested by their own teachers within their entire natural groups, as a classroom activity on thinking assessment. The database was cleaned taking into account responses that were potentially biased, highly empty or lacking attention (as reported by teachers).

Instrument

The Challenges of Thinking Test (CoTT_PE6) is designed to measure six CT skills (Manassero-Mas & Vázquez-Alonso, 2019): prediction and logical reasoning (from the reasoning dimension), comparison (creativity dimension), classification (evaluation dimension), and decision-making and problem-solving (complex processes dimension). Classification assesses the ability to group or separate different elements according to the appraisal of various common or differential features. Prediction and Comparison assess the ability to verify a conclusion from inductive reasoning or from the creative contrast between several statements, respectively. Decision-making and Problem-solving measure the ability to identify the best (worst) decisions/solutions in a particular situation. Logical reasoning assesses simple (simple syllogism) and complex (several pieces of information and conclusions involved simultaneously) deductive abilities.

The skills were agreed by the researchers and the teachers of the participating schools considering their fit to the usual cognitive demands in the sixth grade of primary school (PE6). The researchers selected the test items from the five-skill test previously validated and the piloted items of the item bank through the following criteria: simplicity of wording, ease reading comprehension, balance between item cognitive demand and target students' cognitive development, and motivating and interesting challenge for students (e.g., a simple story on futurist planet exploration, many items with figurative contents and logical reasoning on pencils, books and colors). Then, the researchers assigned each item to the skill that presented the greatest congruence with the item content (Table 1).

Initially, many CT items were selected, translated and adapted from some original standardized CT tests (Ennis & Millman, 2005a, 2005b; Halpern, 2010) and from scholar CT publications (https://www.criticalthinking.com) affordable for

Table 1
Specifications of the two tests applied (CoTT_PE6) in this study to evaluate thinking skills in sixth-grade primary education PE6.

Thinking skills	Item Source	Type	Number of items	
			Initial (48)	Final (31)
Prediction (PREDIC)	Ennis & Millman, 2005a	Verbal	9	6
Comparison (COMPA)		Verbal	7	3
Classification (CLASSIF)	Author elaboration*	Figurative	6	5
Problem-solving (PROBL)	Halpern (2010)	Verbal	6	4
	Author elaboration*	Figurative	4	1
Decision-making (DECISION)	Author elaboration*	Mixed	9	5
Logical reasoning (LOG-REAS)	Author elaboration*	Figurative	-	2
	Ennis & Millman, 2005b	Verbal	7	5

Note. * Translated and adapted from open materials (https://ww.criticalthinking.com).

primary students. The publications provided the figurative items and both made explicit the cognitive demands and the specific item skills. Further, the researchers' professional judgment consensually scrutinized and selected a bunch of items again, by reviewing the best fit between item content and skills and between item's cognitive demand and primary students' cognitive stage. The subsequent item set was piloted and the analysis of the results set up the 48-item CoTT, which is analyzed here (Table 1). All in all, the scholarly nature of the managed item sources and the selection and pilot processes warrant an accurate item-skill correspondence and a sound contribution to assure the content validity of CoTT *ab initio*.

The items pose authentic and motivating thinking challenges for students through a variety of scenarios and situations, where information is communicated by several means of representation (verbal, numerical, and figurative), and their cognitive demands fit the represented skill and the students' evolutive stage. Further, the item contents are independent of the school curricula (for example, they do not involve numerical calculations), so achieving the correct answer just requires applying reasoning to the existing information and does not need any previous knowledge. Therefore, CoTT is considered a culture-free test, as its challenges are not mediated by academic knowledge, as is often the case. For example, the Science CT test requires primary science curriculum knowledge to answer correctly (Mapeala & Siew, 2015).

The response formats are mostly closed and the four items asking for a short open answer were dealt with a simple rubric to code them as correct/wrong. This format allows for a standardized, fast, valid, and reliable evaluation of thinking skills, for the development of diagnostic baselines to objectively compare different research, programs, and teaching methodologies, and for practical use by teachers. Correct answers scored one point and incorrect answers zero, the score of each skill is the sum of the correct answers achieved in their assigned items, and the overall score is the sum of the total correct answers, which is considered an estimate of the students' global CT performance.

Data collection and analysis

The CoTT was applied to the students by their teachers within their class group, as an ordinary regular activity of school evaluation to stimulate the students' efforts and motivation on thinking. To ensure the application consistency the authors supervised the class applications, which followed the usual standardized guidelines for tests, without any time limit (usually a class period); the guidelines were written at the first screen of the test digital device and were read aloud by teachers and students.

The procedures involved a two-stage action. The first stage performed the construction of the 48-item CoTT on the basis of the CT item bank, its application to a large sample of sixth-grade primary school students and the analysis of their responses through exploratory (EFA) and confirmatory factor analysis (CFA) of the restricted 46-item CoTT_PE6_46 (two items of a triplet were eliminated). The second stage involved analyzing the former results to eliminate some items with inadequate psychometric traits to leave a final shorter 31-item form CoTT_PE6_31, which is again scrutinized through EFA and CFA to set up its psychometric properties.

Data were processed with the programs SPSS (25), Amos 23.0.0 and Factor (12.01.02). SPSS and Amos are well-known statistical tools and Factor provides computation of tetrachoric correlations (appropriate for test dichotomous scores) and develops EFAs and CFAs that extract factors with a robust method of unweighted least squares (RULS), parallel analysis, bootstrap sampling, Promin rotations and several indices of reliability (Ferrando & Lorenzo-Seva, 2017, 2018; Lorenzo-Seva & Ferrando, 2019).

Results

The statistical descriptors of the items of the two tests in the two stages of the study, obtained from the students' answers, are summarized in Table 2.

The second column of Table 2 presents the correct answer average for the 48 items that constitute the starting point of this study. Most of the items (39) achieve an intermediate mean of correct answers (.30 – .70), a few items (3) are very easy (*M* > .70), and others (6) are difficult (*M* < .30). The item distribution by quartiles is as follows: 12 (25.0%) items are in the lower quartile (1), 9 (18.7%) items are in the lower-middle quartile (2), 13 (27.1%) items are in the upper middle quartile (3) and 14 (29.2%) items are in the upper quartile (4). The overall average of correct answers is close to 50% (*M* = .485), which confirms the moderate difficulty of the test, as befits this kind of test.

Lastly, three items of the initial 48-item CoTT_PE6 instrument displayed quite high correlations among them to consider they form a triplet. Thus, two of them (PROBL11 and PROBL12) were eliminated, and the subsequent analyses refer to the remaining 46 items (Ferrando et al., 2022).

Factor analysis

The analysis of the 46 items with the RULS method and tetrachoric correlations have got an unfavorable value for the Kaiser-Meyer-Olkin (KMO) parameter (0.206). However, a solution of six empirical factors (as theoretically required) produced quite acceptable goodness-of-fit parameters through minimum rank parallel CFA (Table 3). However, the rotated six-factor solution did not allow a theoretically consistent interpretation of the factors, as it displayed many items with low factor loads or with overlaps on several factors.

To increase the model coherence, the 46 items were scrutinized to remove the items that attain most of the following psychometric parame-

Table 2
Proportion of item average correct answers (difficulty index) of the CoTT_PE6 instrument in a sample of 6th-grade students (*n* = 655).

Initial items (48)		Average Correct Answers (0-1)	Standard deviation	Quartile	Final items (31)
V1	PREDIC1	.623	.485	4	PREDIC1
V2	PREDIC2	.431	.496	2	*
V3	PREDIC3	.499	.500	3	PREDIC3
V4	PREDIC4	.400	.490	2	*
V5	PREDIC5	.791	.407	4	PREDIC5
V6	PREDIC6	.736	.441	4	PREDIC6
V7	PREDIC7	.708	.455	4	PREDIC7
V8	PREDIC8	.380	.486	2	PREDIC8
V9	PREDIC9	.638	.481	4	*
V10	COMPA1	.441	.497	2	COMPA1
V11	COMPA2	.565	.496	3	*
V12	COMPA3	.431	.496	2	*
V13	COMPA4	.521	.500	3	*
V14	COMPA5	.499	.500	3	COMPA5
V15	COMPA6	.501	.500	3	COMPA6
V16	COMPA7	.356	.479	1	*
V17	CLASSIF1	.635	.482	4	PROBLº
V18	CLASSIF2	.553	.498	3	CLASSIF2
V19	CLASSIF3	.559	.497	3	CLASSIF3
V20	CLASSIF4	.653	.476	4	CLASSIF4
V21	CLASSIF5	.663	.473	4	CLASSIF5
V22	CLASSIF6	.640	.480	4	CLASSIF6
V23	PROBL1	.649	.478	4	*
V24	PROBL2	.562	.497	3	PROBL2
V25	PROBL3	.485	.500	3	*
V26	PROBL4	.325	.469	1	PROBL4
V27	PROBL5	.627	.484	4	PROBL5
V28	PROBL6	.621	.485	4	PROBL6
V29	DECISION1	.351	.478	1	*
V30	DECISION2	.282	.451	1	DECISION2
V31	DECISION3	.298	.458	1	*
V32	DECISION4	.328	.470	1	DECISION4
V33	DECISION5	.426	.495	2	DECISION5
V34	DECISION6	.185	.388	1	DECISION6
V35	DECISION7	.145	.352	1	*
V36	DECISION8	.646	.479	4	*
V37	DECISION9	.464	.499	2	DECISION9
V38	PROBL9	.211	.408	1	LOG-REASº
V39	PROBL10	.426	.495	2	PROBL10
-	PROBL11	.536	.499	3	-
-	PROBL12	.377	.485	2	-
V40	LOG-REAS1	.540	.499	3	LOG-REAS1
V41	LOG-REAS2	.554	.497	3	LOG-REAS2
V42	LOG-REAS3	.305	.461	1	*
V43	LOG-REAS4	.588	.493	4	LOG-REAS4
V44	LOG-REAS5	.235	.424	1	*
V45	LOG-REAS6	.574	.495	3	LOG-REAS6
V46	LOG-REAS7	.327	.469	1	LOG-REAS7

Note. - Items eliminated for being part of a triplet (observed correlations 0.853; 0.957; 0.896).
* Items eliminated in the validation process of the final instrument.
º New skill of the items that have changed its final skill assignment through the validation process.

ters: negative values in the standardized tetrachoric correlation matrix; sampling adequacy measure values less than .50; negative, null or crossed factor loads in the rotated matrix; low standardized regression loads between the item and the empirical factors; MIREAL (Mean of Item REsidual Absolute Loadings) parameter greater than .30; approximately meeting the optimal standard that 75% items of the item pool achieve an intermediate range (.40 – .60) of the relative difficulty index, and the remaining are evenly distributed in both tails (Ferrando & Lorenzo-Seva, 2017).

The joint qualitative application of the former criteria leads to the identification of 12 items that showed deficiencies in several criteria, thus deciding their elimination to improve the test (PREDIC2, PREDIC4, PREDIC9, COMPA3, COMPA4, PROBL1, DECISION1, DECISION3, DECISION8, LOG-REAS3, LOG-REAS5, COMPA7).

The resulting set of 34 items was again analysed with the RULS method and tetrachoric correlations. The results showed a better but still low value of the KMO parameter (0.545), although other parameters improved, and some were excellent. The average difficulty index was now .503, and the reliability factor was high (α = .83). The six empirical factors explained 46% of the variance, and the excellent goodness-of-fit indices showed that the data adequately fit an empirical six-factor model (RMSEA= .036 CFI = .973, GFI = .956, RMSR = .05), which also shows closeness to unidimensionality assessment (MIREAL = .193) (Ferrando & Lorenzo-Seva, 2018). Despite these good parameters, three items presented zero or slightly negative factor loadings on all the factors of the rotated matrix and nonsignificant standardized regression coefficients. Further, two of them also displayed multiple negative correlations, and still another item showed a high difficulty index. Therefore, these three items were also removed (COMPA2, PROBL3, DECISION7).

The set of the remaining 31 items was again reanalyzed to get the new values of CFA parameters that may validate the test final form (second column, Table 3), although the KMO parameter (.578) is still moderate. The average difficulty index (.514) and the reliability coefficient are also good (α = .838). The scree-plot displays a main eigenvalue that explains 19% of the total explained variance and a main elbow respect the following eigenvalues; then, the soft decrease displays a smaller elbow between the eigenvalues 6 and 7, where the first six make contributions to the variance well over 5%, and the whole six first factors explain 49%, while the following decreasingly contribute less than 4%. Thus, a CFA was applied to the six-factor model, whose rotated loading matrix showed no cross-factor loads greater than .30 between the factors, only two factor loadings were less than .20 (COMPA1 and LOG-REAS7) and an interpretable structure of factors. The CFA goodness-of-fit indices show an excellent fit for the six-factor empirical model (MIREAL=.186, RMSEA = .032, CFI = .982, GFI = .965, RMSR = .049). Further, the CFA analysis of the six-factor empirical model is close to one-dimension and the reliability parameters (ORION) of factors are also good (.735 – . 999).

To test whether the covariance structure of 31 items can also be satisfactorily explained by a single general factor, a general factor model was analyzed through CFA (right column, Table 3). The comparison of six-factor and general factor models through the likelihood ratio test (chi^2 = 1391.096, *df* = 140, *p* < .000) is significant, which means both models are different. Further, the six-factor model significantly increments its scores of RMSEA, NNFI, CFI, GFI and residuals in relation to those of the general factor model, and attains the thresholds prescribed for these parameters, while the general factor model does

Table 3
Robust statistical parameters of the confirmatory goodness of fit of the contrasted factor models for the initial test (46 items) and the final test (31 items).

	Contrasted models		
	46 items		**31 items**
Extracted factors	6	6	1
Kaiser-Meyer-Olkin	.206	.578	.578
Bartlett (Sig.)	-	.000	.000
Explained variance	.382	.489	.192
Goodness of fit			
RMSEA*	.041	.032	.071
Chi-square	1633.367	489.959	1880.985
Chi-square (p)	.000	.000 (*df* = 294)	.000 (*df* = 434)
NNFI**	.934	.971	.857
CFI***	.951	.982	.867
GFI****	.932	.965	.866
Residuals			
RMSR[o]	.058	.049	.097
WRMSR[oo]	.036	.030	.053
Reliability			
EAP-GLB[ooo]	.981	.963	.963
Omega	.850	.822	.822
Cronbach Alpha	.853	.838	.838
ORION[a]-Factor1	.913 (CLASSIF)	.922 (CLASSIF)	
ORION[a]-Factor2	.809 (PROBL)	.999 (PREDIC)	
ORION[a]-Factor3	.859 (DECIS)	.815 (COMPA)	
ORION[a]-Factor4	.738 (COMPA)	.845 (REAS)	
ORION[a]-Factor5	.857 (REAS)	.741 (DECIS)	
ORION[a]-Factor6	.860 (PREDIC)	.735 (PROBL)	

Note.* Root Mean Square Error of Approximation.
** Normed Fit Index.
*** Comparative Fit Index.
**** Goodness of Fit Index .
[o] Root Mean Square of Residuals (acceptable close to .048).
[oo] Weighted Root Mean Square of Residuals (acceptable fit < 1.0).
[ooo] Expected a Posteriori (EAP) Greatest Lower Bound (GLB) for reliability.
[a] Overall Reliability of fully-Informative prior Oblique N-Expected a Posteriori scores.

not (Calderón-Garrido et al., 2019). These results point out the six-factor model represents the data better than the general factor model.

As the six-factor empirical model identified corresponds to a structure whose constituent elements allow a reasonable interpretation of the model according to the theoretical proposal presented at the beginning of this study, the same names of the factors were retained, namely Classification, Prediction, Comparison, Reasoning, Decision, and Problem. The results of the previous analysis eliminated 17 initial items, due to the detection of some empirical dysfunctions, thus decreasing sharply the number of items that form the empirical factors that are retained for the final form of the test; however, only two items switched their theoretically factor assigned initially, as a consequence of the CFA validation of the final empirical factors. For instance, the CLASSIF1 item was initially and theoretically assigned to the scale Classification and was empirically allocated into the Problem final factor (CLASSIF1_PROBL); the PROBL9 item was initially and theoretically assigned to the scale Problem, yet it was empirically allocated into the final empirical factor Reasoning (PROBL9_REAS). The mixed denominations of these elements, which include both the initial theoretical dimension and their final empirical factor, try to reflect their switched situation (Table 4).

Figure 1 represents the structural equation model corresponding to the loading matrix of Table 4 and the excellent parameters of the CFA (Table 3). The diagram shows the standardized regression coefficients among the latent variable and with the observable variables of the instrument, as well as the proportions of empirically explained variance for each variable. The model depicts strong relationships of five latent scales (Problem solving achieves the highest standardized coefficient) and also suggests some weakness of the Comparison latent scale, probably due to the drastic reduction of its length to just three items.

The model incorporates four residuals correlations that were added because of their high modification indices and the gain in the model fit, as the model without the residual correlations attained worst fit parameters (e.g., higher Chi-square = 677.915, NNFI and CFI < .90) than those reported for the final model (Table 3). Further, the residual correlations may also have some basis due to theoretical similarities of the items.

Analysis of the closeness to the single-dimension of the factors

Each of the six groups of items that make up the six empirical factors obtained from the previous CFAs for the 31-item test (Table 4) were submitted separately to a confirmatory RULS analysis to verify their one-dimensional nature. The overall results obtained in the six factors show adequate goodness-of-fit indices, explained variance and reliability, but also suggest some improvements (Table 5).

Parallel analyses with the RULS method and tetrachoric correlations based on the minimum rank factor analyses confirmed a single-dimension model for the six factors, because the MIREAL parameter presents acceptable values (< .30), with a moderate exception in Decision. These results allow us to consider these six factors as one-dimensional, and, consequently, justify that their scores validly and reliably measure each of the skills operationalized by the items that make up the empirical factors.

The proportion of variance explained by each of the six unique factors is high (.57 –.39), and both reliability values, omega (.873 –.638) and Cronbach's alpha (.870 –.624) are good. Although four factors display good KMO values (.823 –.658), the Prediction and Comparison factors show low KMO values (.535 –.501). Almost all the loadings of the constituent items of empirical factors reach scores greater than .30, with the sole exception of COMPA1 item, which may be the source of the problems of this factor, together with the small number of items that form it (3). The goodness-of-fit parameters of the CFA show that the data obtained for the six factors adequately fit the one-dimension structure as their scores for the six factors are excellent: RMSEA (.03 – .09), CFI (.954 – . 996), GFI (.969 – .997), and RMSR (.019 – . 098).

Empirical analysis of the test validity

The process of CoTT_EP6 construction stemmed from scholar and credible item sourc-

Table 4
Factor loading matrix of the reduced CoTT_PE6_31 test (31 items; Promin rotation).

Variables			**Empirical**	**Factors**		
	Classification	Prediction	Comparison	Reasoning	Decision	Problem
PREDIC1		.415				
PREDIC3		.369				
PREDIC5		1.014				
PREDIC6		.515				
PREDIC7		.339				
PREDIC8		.274				
COMPA1			.126			
COMPA5			.810			
COMPA6			.830			
CLASSIF2	.974					
CLASSIF3	.737					
CLASSIF4	.601					
CLASSIF5	.609					
CLASSIF6	.861					
CLAS1_PROBL						.268
PROBL2						.666
PROBL4						.633
PROBL5						.560
PROBL6						.440
PROBL10						.276
DECISION2					.436	
DECISION4					.440	
DECISION5					.327	
DECISION6					.622	
DECISION9					.757	
PROBL9_REAS				.233		
LOG-REAS1				.805		
LOG-REAS2				.677		
LOG-REAS4				.821		
LOG-REAS6				.570		
LOG-REAS7				.154		
Number of items	5	6	3	6	5	6
Reliability (ORION[a])	.922	.999	.815	.845	.741	.735
Explained variance	.192	.258	.319	.378	.434	.489

Note. Loadings below .30 were eliminated (except for six loadings that correspond to items theoretically assigned to a factor).
[a] Overall Reliability of fully-Informative prior Oblique N-Expected a Posteriori scores.

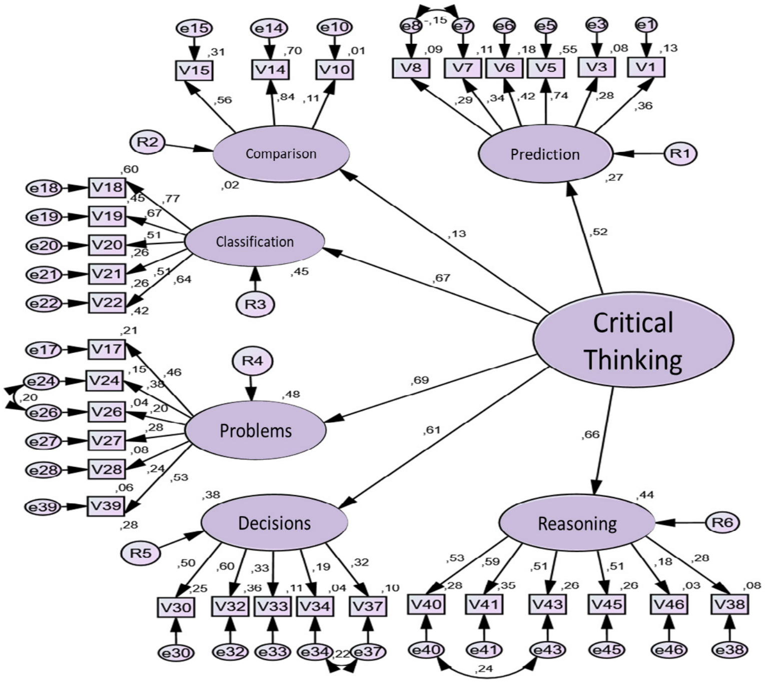

Figure 1
Diagram of structural equations corresponding to the CoTT_PE6_31 instrument.

Table 5
Statistical parameters of the robust goodness of fit of confirmatory factor analysis (CFA) for the closeness to single-dimensional model of each of the six empirical scales resulting from the factorization of the CoTT_PE6_31 instrument.

CFA Statistics	**Prediction**		**Comparison**		**Classification**		**Problem**		**Decision**		**Reasoning**	
	Item	Load	Item	Load	Item	Load	Item	Load	Item	Load	Item	Load
	PRE1	.457	COM1	.133	CLA2	.915	CLA1	.310	DEC2	.585	PRO9	.355
	PRE3	.378	COM5	1.000	CLA3	.766	PRO2	.627	DEC4	.638	REAS1	.741
	PRE5	1.000	COM6	.664	CLA4	.645	PRO4	.645	DEC5	.394	REAS2	.685
	PRE6	.534			CLA5	.658	PRO5	.556	DEC6	.440	REAS4	.744
	PRE7	.346			CLA6	.807	PRO6	.385	DEC9	.583	REAS6	.599
	PRE8	.333					PRO10	.411			REAS7	.194
Kaiser-Meyer-Olkin	.535		.501		.823		.700		.658		.753	
Bartlett (P-Sig.)	.000		.000		.000		.000		.000		.000	
Explained variance	.390		.567		.661		.408		.425		.404	
One-dimensionality												
MIREAL*	.297		.293		.229		.250		.330		.215	
Goodness of fit												
RMSEA**	.061		-		.051		.040		.091		.031	
RM Ji Squared	3.831		-		13.625		1.221		31.973		22.943	
Ji Square(P)	.000		-		.197		.071		.000		.064	
NNFI ***	.956		-		.992		.978		.908		.990	
CFI****	.974		-		.996		.989		.954		.993	

CFA Statistics	Prediction	Comparison	Classification	Problem	Decision	Reasoning
GFI*****	.970	.997	.996	.991	.972	.969
Residuals						
RMSRº	.055	.098	.055	.045	.019	.074
WRMSRºº	.038	.063	.039	.041	.012	.043
Reliability						
EAP-GLBººº	.792	.696	.902	.686	.750	.812
Omega	.670	.679	.873	.638	.663	.743
Cronbach Alpha	.658	.555	.870	.624	.656	.730

Note.* Mean of Item Residual Absolute Loadings (unidimensional < .30).
** Root Mean Square Error of Approximation.
*** Normed Fit Index.
**** Comparative Fit Index.
***** Goodness of Fit Index.
º Root Mean Square of Residuals (acceptable model if close to .048).
ºº Weighted Root Mean Square of Residuals (acceptable fit < 1.0).
ººº Expected a Posteriori (EAP) Greatest Lower Bound (GLB) for reliability.

es that provided quality and affordable CT items. The items were scrutinized, analyzed, selected and piloted by researchers, in order to test their cognitive demand to PE6 students and prepare the elaboration of the 48-item CoTT_PE6. The reliability of the sources and the selection processes warrant some ab initio content validity of CoTT through sound fitness between item contents, skills, cognitive demands and students' abilities. This section aims to further develop the empirical CoTT validity.

Parallel tests are often used by researchers to trial validity through external criteria. However, the scarcity of CT tests for primary students, which mainly motivates this study, makes this way impractical. Instead, each of the six theoretical skills that conform to CoTT are taken as external criteria for the remaining skills, so that the analysis of skill intercorrelations develops a validity test. Regardless of the debate about the importance of the general or specific context of CT education, the most widespread argument in favor of CT is its impact on learning (e.g., O'Hare & McGuinness, 2015). Thus, learning is operationalized here through students' subject school grades at the end of the school year, which are numerically assessed by teachers (1-10) and used here as an external criterion to test the CoTT_PE6 validity. Finally, the correlations and variances of the empirical factors are examined to add evidence on CoTT validity.

Correlations among critical thinking skills

The descriptive statistics and the correlations among the six CoTT theoretical skills are displayed and analyzed here assuming that all of them are significant and positive, considering that all of them measure an aspect of the CT construct (Table 6). All six skills display cases attaining their maximum and minimum scores, which display the range (9-43 and 4-31) for the total score of the two CoTT forms. The mean scores of skills show that Classification and Prediction tend to display the highest mean score, while Decision and Logical Reasoning have got the lowest scores. The asymmetry and kurtosis (not shown in Table 6) have got normal scores for the six skills and the total score.

Table 6

Descriptive statistics of the six theoretical skills (top) emerging from the empirical factorization of CoTT_PE6_46 and CoTT_PE6_31 (*n* = 655) and their Pearson correlation coefficients (bottom), where the upper triangle corresponds to the skills of CoTT_PE6_31 (columns) and the lower triangle to the CoTT_PE6_46 skills (rows).

	Descriptive Statistics of CoTT_PE6_46						
	Prediction9	Comparison7	Classification6	Problem8	Decision9	Log-Reason7	Total46
Range	9	7	6	8	9	7	34
Minimum	0	0	0	0	0	0	9
Maximum	9	7	6	8	9	7	43
Mean	5.206	3.313	3.702	3.907	3.125	3.124	22.377
Error std.	0.075	0.055	0.074	0.065	0.072	0.064	0.251
Std. deviation	1.924	1.417	1.889	1.672	1.854	1.631	6.416

	Descriptive Statistics of CoTT_PE6_31						
	Prediction6	Comparison3	Classification5	Problem6	Decision5	Log-Reason6	Total31
Range	6	3	5	6	5	6	27
Minimum	0	0	0	0	0	0	4
Maximum	6	3	5	6	5	6	31
Mean	3.737	1.441	3.067	3.197	1.686	2.794	15.922
Error std.	0.058	0.040	0.068	0.061	0.052	0.064	0.198
Std. deviation	1.474	1.024	1.728	1.573	1.334	1.641	5.056

CoTT_PE6_46 Skills	**CoTT_PE6_31 Skills**					
	Prediction6	Comparison3	Classification5	Problem6	Decision5	Log-Reason6
Prediction9		.063	.205	.265	.206	.142
Comparison7	.220**		.235	.190	.031	.384**
Classification6	.362**	.213**		.377**	.275*	.037
Problem8	.182**	.143**	.300**		.340*	.341*
Decision9	.282**	.162**	.373**	.263**		.060
Log-Reas7	.282**	.209**	.286**	.252**	.216**	

Note.* Correlation significant at the .05 level (bilateral).
** Correlation significant at the .01 level (bilateral).

All Pearson correlations between the theoretical skills are positive and significant (p < .01) for the CoTT_PE6_46. The CoTT_PE6_31 inter-skill correlations display lower scores than the former, though all them are still positive (Table 6). Thus, the inter-skills correlations are positive and mostly significant as expected to justify the internal validity of CoTT. The correlations among the empirical factors of both CoTT forms are overall higher than the correlations between the theoretical sub-scales (Table 6), which may justify much better than the correlations of table 6 both, the factors' high reliability and the relative weakness of the Comparison factor (correlations close to zero).

Correlations with school subject grades as external criterion

The empirical analysis of CoTT validity in regard to external criteria (school grades) applies correlational methods. To this aim, an incidental subsample of participants (*n* = 52), those whose final-course grades were available, is used. In spite of the size, the subsample is diverse, as it comes from three schools (two public and one public-funded), which are located at a small town and at the center and the periphery of a large city. Table 7 displays the descriptive statistics and Pearson correlations of the school grades and skills.

The descriptive statistics of grades show that the distribution of students' grades is quite homogeneous across subjects, and the asymmetry and kurtosis scores stay within acceptable ranges (not shown). It is worth highlighting that only Natural Science, Catalan Language and Math display 4 as minimum grade, which means that some sixth-graders have got negative final grades (under 5). Further, Physical Education showed the highest average grade and the minimum standard deviation, whilst Catalan Language displays the lowest average grade (Table 8).

Table 7
Pearson intercorrelations among school subject grades (top of the table) and the descriptive statistics of grades (bottom) for the incidental subsample (*n* = 52).

Subjects	Natural Sc.	Social Sc.	Catalan L	Spanish L.	Art Ed.	Physical Ed.	Math	Religion-Values	English L.
Natural Sciences	-	.858**	.819**	.840**	.708**	.544**	.733**	.527**	.742**
Social Sciences		-	.857**	.863**	.559**	.449**	.700**	.700**	.648**
Catalan Language			-	.899**	.590**	.466**	.794**	.696**	.681**
Spanish Language				-	.637**	.470**	.796**	.723**	.768**
Art Education					-	.630**	.541**	.343*	.693**
Physical Education						-	.394**	.344*	.499**
Mathematics							-	.665**	.694**
Religion-Values								-	.545**
English Language									-
	Descriptive statistics of grades (range of grade scores 1-10; grades under 5 are negative)								
Range	6	5	5	5	5	5	6	5	5
Minimum	4	5	4	5	5	5	4	5	5
Maximum	10	10	9	10	10	10	10	10	10
Mean	7.58	7.92	7.31	7.46	8.21	8.5	7.92	8.31	8
Std. deviation	1.719	1.453	1.515	1.578	1.637	1.111	1.747	1.489	1.521

Table 8
Pearson intercorrelations between school subject grades and CT theoretical skills of the two forms of CoTT for the incidental subsample (*n* = 52). The correlations in bold pinpoint the only skill that significantly enters the subject grade prediction model of the lineal regression analysis.

CoTT Skills	**Correlations**								
	Natural Sc.	Social Sc.	Catalan L	Spanish L.	Art Ed.	Physical Ed.	Math	Religion-Values	English L.
CoTT_PE6_46									
Prediction9	.225	**.319***	.236	**.392****	.022	.02	.265	.255	.243
Comparison7	.228	.182	.255	.316*	.145	.121	.323*	.16	**.412****
Classification6	.177	.055	.08	.241	**.306***	.059	.259	.023	.223
Problem8	.219	.128	**.300***	.332*	.219	.203	**.479****	.213	.256
Decision9	.164	.13	.225	.271	.23	.251	.264	.061	.117
Log-Reas7	.032	.06	.158	.121	-.039	-.071	.249	.203	.207
Total46	.294*	.247	.356**	.476**	.249	.161	.529**	.265	.408**
Explained variance(%)***	8.6	10.2	12.7	22.7	9.4	2.6	28.0	7.0	16.6
CoTT_PE6_31									
Prediction6	.113	.179	.107	.255	-.047	-.048	.156	.127	.140
Comparison3	-.042	-.150	-.034	-.055	.061	.125	-.051	-.236	.056
Classification5	.146	-.003	.010	.192	**.323***	.109	.216	-.075	.200
Problem6	**.329****	.257	**.415****	**.457****	.224	.103	**.520****	**.361****	**.321***
Decision5	.082	.162	.256	.247	.176	**.291***	.235	.155	.058
Log-Reas6	-.035	-.053	.051	.052	-.107	-.246	.185	.060	.135
Total31	.182	.122	.236	.355**	.177	.056	.397**	.134*	.287
Explained variance***	10.8	6.6	17.2	20.9	10.4	8.5	27.0	13.0	10.3

Note. * Correlation significant at the .05 level (bilateral).
** Correlation significant at the .01 level (bilateral).
*** Shared variance between subject grades and skills (computed through lineal regression analysis of grades and CT-skills as the square of the subject's bold correlation coefficient, due to the single-skill prediction model obtained).

The correlations between subjects are all positive and significant, and it is worth noting the highest correlations between Catalan, Spanish, Social and Natural Sciences, whilst Physical Education and Art Education display the lowest correlations with the others.

The correlations between CT skills and school grades are mainly positive as expected, although only a few are statistically significant. The total score of both CoTT forms correlates positively with all subjects, which means CoTT total scores predict grades and thus support the predictive validity of both forms and inform the amount of explained variance for each grade (top 28% for Math). From the point of view of the specific skills, Problem-solving displays the highest correlations with the subject grades, and significantly correlates with three/six subjects (depending on the form) for the two forms. At the opposite extreme, Decision and Logical Reasoning tend to display the lowest set of correlations with subjects. Problem solving of CoTT_PE6_31 largely displays the highest correlations with almost all subjects, while CoTT_PE6_46 displays a wider distributed pattern of skills than CoTT_PE6_46.

Further, the correlations show some specific correlation patterns of the subjects for both forms of CoTT. Overall, CoTT_PE6_46 tends to display higher correlations than CoTT_PE6_31, where a few correlations are negative yet non-significant. The highest and significant correlations ($p < .01$) correspond to Mathematics (maximum) and the language subjects (Catalan, Spanish and English), whereas the lowest (non-significant) correlations correspond to Physical Education (minimum). From the perspective of the subjects, the leading correlation of each subject makes sense of the subject curriculum; for instance, Problem Solving is the top correlated skill for Mathematics and Natural Sciences, where problem solving is a core learning activity. Again, Decision Making is the top correlated skill for Physical Education, where sports continuously practice decision making, Classification (mainly made of figurative items) is top for Art and Prediction (causes and consequences) for Social. Comparison and Problem Solving are top for language subjects. This association between subjects and its leading skills also put forward a qualitative predictive validity of CoTT.

Finally, in order to discriminate the predictive power of the different single skills (predictors) on each subject grade, a forward stepwise lineal regression analysis was performed. Of course, the subjects lacking significant correlations have not got significant predictors and the subjects that display only one significant skill correlation this skill is the only predictor. However, the subjects having two or more significant skill correlations have got a prediction equation with only one predictor too (the highest correlation). The predictors of each subject are bold in Table 8 and this qualitative association between subjects and one specific skill also adds to the predictive validity of CoTT.

All in all, the correlations between CT skills and school grades are mainly positive as expected, which suggest the predictive validity of the CoTT in front of an external criterion as school subject grades. The above correlational profiles suggest specific trendy associations between subjects and skills, which could be rationally justified. However, these trends should be better elucidated through large samples and empirical CFA.

Discriminant validity of the model

The discriminant validity of the factor model was verified through the application of the Fornell-Larcker criterion, which requires the average factor variances to be higher than the correlations of each factor with the other factors (Fornell & Larcker, 1981). The average factor variances are computed from the data of Figure 1 and the comparison with the inter-factor correlations show that the Fornell-Larcker criterion is satisfied by the model. Thus, the model's discriminant validity is confirmed.

In sum, the positive and significant inter-skill correlations support the concurrent validity (relatively higher correlations across skills, as they all belong to the CT construct). Then, the computed variances of the factors are higher than the inter-correlations and confirm the discriminant validity of CoTT_EP6. Further, the correlations between CT skills and a theoretically-related external criterion (school subject grades) confirm that both constructs are positively and significantly correlated, where the subject correlations with the total CT score are especially high, also underlining the transversal importance of CT for school learning. Thus, the above correlational analyses support the validity of the CoTT_EP6 test.

Discussion

This study provides evidence about the validity and reliability of the Challenges of Thinking (CoTT_PE6) instrument to assess CT, a culture-free test (independent of the school curriculum) that is adapted to the evolutive and learning stage of sixth graders (11 year old). The contributions of CoTT arise from the inner transversal impact of CT on learning and the increasing extension of CT teaching in schools, with the consequent need to evaluate the educational results, as well as the lack of CT assessment instruments, which are adapted to younger students and appropriate for use in the classroom (Aktoprak & Hursen, 2022; Ennis, 2009).

The study follows the general prescriptions of test development to establish the psychometric properties of CoTT_PE6 (Ferrando et al., 2022; Muñiz & Fonseca-Pedrero, 2019). The study validates a six-factor empirical model of the final 31-item CoTT, which confirms a parsimonious and coherent interpretation of the theoretical factor structure postulated for CoTT_PE6_31 (Prediction, Comparison, Classification, Problem solving, Decision making and Logical Reasoning). The CFA goodness-of-fit parameters for the six-factor model are excellent: Chi-square (489.959, p = .000), RMSEA (.032), NNFI (.971), CFI (.982), GFI (.965), and RMSR (.049). In addition, the reliability indices of the whole CoTT_PE6_31 (.838) and each of the six identified empirical factors reach good scores (ORION: .999, .815, .922, .735, .741, .845), following the order of factors of the previous paragraph. The one-dimension structure for each of the six factors of the model is also supported by their CFA individual goodness-of-fit parameters, so they can be independently used in measurements. Their individual reliability is also acceptable (alpha: .658, .555, .870, .624, .656 and .730), despite some lower values, possibly due to the structural effect of shortening the length of each factor.

Validation evidence has been widely displayed along the results section through several confirmatory milestones, such as the credibility of the scholar sources that provided the starting bank of items, the previous piloting of many items that lead to construct the first CoTT form, and the correlational validity tests that were performed through external criteria (school subject grades), internal criteria (correlations of different factors of CoTT) and the computation of factor variances and. All in all, the CoTT validity results through the predictive validity of CT skills on grades, as well as the higher average variance of factors than the correlations among factors, advocate the claims for the transversal relevance of CT in regard of learning (European Union, 2014; OECD, 2018; UNESCO, 2016).

The psychometric validation of the CoTT_PE6, together with its simplicity of application and scoring, endorses its direct and practical application in primary education: this useful and functional tool makes thinking and its progress visible in primary classrooms and educational research. Educators and researchers can easily diagnose and evaluate CT to test the effectiveness of CT intervention programs (Colom et al., 2014; Saiz, 2017). In addition, CoTT_PE6 allows monitoring the progress of CT skills in longitudinal studies of the educational system, so that it can assess the impact of skills on learning and vice versa, which are crucial aspects of the quality of education (Hattie, 2012; OECD, 2018; UNESCO, 2016).

The CoTT_PE6 instrument also has some limitations that arise from its design. The first limitation is the obvious restriction to the six skills it contains, as they are considered appropriate skills for primary education. Overall, the validation process displays some limitations and suggests some corresponding future actions, such as performing test-retest reliability or increasing the sample of the predictive validity through grades. Another limitation is the modest values of some KMO indices, especially for Prediction and Comparison skills, which reflect a tension between opposing psychometric and difficult-to-balance demands; on the one hand, each item must provide differential variance from other items to obtain excellent KMO values; on the other hand, this differential variance partially opposed to the most basic principle of each item, namely, contributing to measuring the common cognitive ability of each skill. Thus, a balanced combination of new items and new samples is required to improve KMO test scores. Moreover, the small number of items that compose the Comparison factor (3) possibly harms the overall goodness of fit, despite its GFI parameter is still excellent (.997), yet a future stronger set of items may address this weakness. Consequently, the refinement of items and applications to new non-convenience samples is expected to overcome these limitations (Ferrando et al., 2022; Muñiz & Fonseca-Pedrero, 2019).

Future application of the CoTT_PE6_31 is expected to provide additional evidence for increasing the instrument's validity and reliability. In particular, the analysis of effects and consequences across different groups of students (gender, age, etc.) and across time (test-retest stability) may provide higher response variability and contribute to improve the test validity and reliability, and to consolidate its educational functionality in aspects such as the standardization in different groups, the relationship with other cognitive measures of CT and school grades, as well as the predictive validity between them, and, in short, to increase the visibility of thinking in education (Lopes et al., 2018).

Conclusions

The 31-item final form of the newer Challenges of Thinking test evaluates six CT skills (Prediction, Comparison, Classification, Problem-Solving, Decision-Making and Logical Reasoning) in 6th graders. This test evaluates six empirical factors that constitute genuine cognitive skills of CT, overlap and fit the theoretical description of CT skills and are independent of the previous knowledge and unidimensional, thus allowing the independent evaluation of each skill, unlike other similar instruments that sometimes mix skills, dispositions and knowledge. Its validation results show that the test is valid, reliable, functional, useful, its response processes are standardized, the internal structure shows quite good CFA parameters and adequate evidence of factor discriminant validity as well as a positive relationship with school grades, as an external variable. Finally, the test application satisfies conditions of time and material economy to easily evaluate, research and make thinking visible in primary education classrooms.

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). Structure of the competences of the XXI century in students of the educational field. Personal influential factors. *Educación XX1, 23*(1), 45-74. https://doi.org/10.5944/educXX1.23853

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies, 31*(3), 285-302. https://www.tandfonline.com/toc/tcus20/31/3?nav=tocList

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

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. (2009). An annotated list of critical thinking tests. https://criticalthinking.net/how-can-critical-thinking-skills-be-tested

Ennis, R. H., & Millman, J. (2005a). *Cornell Critical Thinking Test Level X*. The Critical Thinking Company. https://www.criticalthinking.com/cornell-critical-thinking-tests.html

Ennis, R. H., & Millman, J. (2005b). *Cornell Critical Thinking Test Level Z*. The Critical Thinking Company. https://www.criticalthinking.com/cornell-critical-thinking-tests.html

European Union. (2014). *Key competence development in school education in Europe. KeyCoNet's review of the literature: A summary*. Key Competence Network. 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., Blohm, S. W., Howard, K. L., & Giancarlo, C. A. (1998). *California Critical Thinking Skills Test: Manual (Revised)*. California Academic Press.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*(2), 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*(5), 762-780. https://doi.org/10.1177/0013164417719308

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decalogue for the factor analysis of Test Items. *Psicothema, 34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

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

Fisher, A. (2021). What critical thinking is. In J. A. Blair (Ed.), *Studies in critical thinking* (2nd ed., pp. 7-26). University of Windsor.

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* (Extra). https://revistas.upn.edu.co/index.php/TED/issue/view/583

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

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50. https://doi.org/10.2307/3151312

Calderón-Garrido, C., Navarro-González, D., Lorenzo-Seva, U., & Ferrando-Piera, P. J. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor-analytic approach for assessing the dimensionality of tests and items. *Psicothema, 31*(4), 450-457. https://doi.org/10.7334/PSICOTHEMA2019.153

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 (4th ed.)*. Lawrence Erlbaum.

Halpern, D. F. (2010). *Halpern Critical Thinking Assessment*. SCHUHFRIED. https://www.schuhfried.com

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

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

International Society for Technology Education. (n.d.). ISTE Standards: For Educators. International Society for Technology Education. https://iste.org/standards/educators

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). Critical thinking test for elementary 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*(1), 99-106. https://doi.org/10.24265/liberabit.2019.v25n1.08

Madison, J. (2004). James Madison Critical Thinking Course. The Critical Thinking Company. https://www.criticalthinking.com/james-madison-critical-thinking-course.html

Manassero-Mas, M. A., & Vázquez-Alonso, Á. (2019). Taxonomía de las destrezas de pensamiento: Una herramienta clave para la alfabetización científica. En M. D. Maciel & E. Albrecht (org.), *Ciência, Tecnologia & Sociedade: Ensino, Pesquisa e Formação,* (pp. 17-38). UNICSUL.

Manassero Mas, M. A., & Vázquez Alonso, Á. (2020a). Evaluación de destrezas de pensamiento crítico: Validación de instrumentos libres de cultura. *Tecné, Episteme y Didaxis, 47*, 15-32. https://doi.org/10.17227/ted.num47-9801

Manassero-Mas, M. A., & Vázquez-Alonso, Á. (2020b). Las destrezas de pensamiento y las calificaciones escolares en educación secundaria: Validación de un instrumento de evaluación libre de cultura. *Tecné, Episteme y Didaxis, 48*, 33-54. https://doi.org/10.17227/ted.num48-12375

Manassero-Mas, M. A., Moreno-Salvo, A., & Vázquez-Alonso, Á. (2022). Development of an instrument to assess young people's attitudes toward critical thinking. *Thinking Skills and Creativity, 45*, 101100. https://doi.org/10.1016/J.TSC.2022.101100

Manassero-Mas, M. A., & Vázquez-Alonso, Á. (2023). Evaluación de las destrezas del pensamiento crítico: Un diagnóstico de los estudiantes de primaria. *Revista Evaluar, 23*(2), 40-56. https://doi.org/10.35670/1667-4545.v23.n2.42069

Manassero-Mas, M. A., & Vázquez-Alonso, Á. (2024). Visibilizar las destrezas de pensamiento en educación primaria: Desarrollo psicométrico de un instrumento de evaluación. *Bordón. Revista de Pedagogía, 76*(1), 119-139. https://doi.org/10.13042/BORDON.2024.95702

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

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

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

Organisation for Economic Co-operation and Development [OECD]. (2018). *Future of education and skills.* https://www.oecd.org/en/topics/future-of-education-and-skills.html

O'Hare, L., & McGuinness, C. (2015). The validity of critical thinking tests for predicting degree performance: A longitudinal study. *International Journal of Educational Research, 72*, 162-172. https://doi.org/10.1016/j.ijer.2015.06.004

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. https://www.criticalthinking.org/data/pages/40/fe9f23bd821fcc3ac920a3ce58352412513525db13333.pdf

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

Rivas, S. F., & Saiz, C. (2012). Validation and psychometric properties of the PENCRISAL Critical Thinking Test. *Revista Electrónica de Metodología Aplicada, 17*(1), 18-34. https://reunido.uniovi.es/index.php/Rema/issue/view/766

Saiz, C. (2017). Pensamiento crítico y cambio [*Critical thinking and change*]. Pirámide.

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

Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2013). *Thinking-based learning*. SM.

Tremblay, K. (2013). OECD Assessment of Higher Education Learning Outcomes (AHELO). En Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., Fege, J. (eds.) Modeling and Measuring Competencies in Higher Education. Professional and Vet Learning, vol 1 (pp 113–126). Sense Publishers. https://doi.org/10.1007/978-94-6091-867-4_8

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/ark:/48223/pf0000245656

Valenzuela, J. (2008). Thinking and deep learning skills. *Revista Iberoamericana de Educación, 46*(7), 1-9. https://doi.org/10.35362/rie4671914

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), 219-245. https://doi.org/10.1080/19462166.2015.1123772

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

Whiting, K. (Oct 21, 2021). These are the top 10 job skills of tomorrow – and how long it takes to learn them. World Economic Forum. https://www.weforum.org/agenda/2020/10/top-10-work-skills-of-tomorrow-how-long-it-takes-to-learn-them