education sciences

MDPI

Article

Development and Validation of a Critical Thinking Assessment-Scale Short Form

Rita Payan-Carreira [1,*], Ana Sacau-Fontenla [2], Hugo Rebelo [3], Luis Sebastião [3] and Dimitris Pnevmatikos [4]

[1] CHRC—Compreensive Health Research Centre, Department of Veterinary Medicine, Universidade de Évora Pole at Mitra, 7002-774 Évora, Portugal
[2] Faculdade de Ciências Humanas e Sociais, Universidade Fernando Pessoa (UFP), 4249-004 Porto, Portugal
[3] CIEP—Research Centre in Education and Psychology, Universidade de Évora Pole at Mitra, 7005-345 Évora, Portugal
[4] Department of Primary Education, University of Western Macedonia—UOWM, GR-53100 Florina, Greece
* Correspondence: rtpayan@uevora.pt

Abstract: This study presents and validates the psychometric characteristics of a short form of the Critical Thinking Self-assessment Scale (CTSAS). The original CTSAS was composed of six subscales representing the six components of Facione's conceptualisation of critical thinking. The CTSAS short form kept the same structures and reduced the number of items from 115 in the original version, to 60. The CTSAS short form was tested with a sample of 531 higher education students from five countries (Germany, Greece, Lithuania, Romania, and Portugal) enrolled in different disciplinary fields (Business Informatics, Teacher Education, English as a Foreign Language, Business and Economics, and Veterinary Medicine). The confirmatory analysis was used to test the new instrument reliability, internal consistency, and construct validity. Both the models that hypothesized the six factors to be correlated and to tap into a second-order factor representing the complex concept of critical thinking, had acceptable fit to the data. The instrument showed strong internal consistency ($\alpha = 0.969$) and strong positive correlations between skills and between the skills and the overall scale ($p < 0.05$). Despite the unbalanced sex distribution in the population (close to 75% females), the instrument retained its factorial structure invariance across sexes. Therefore, the new instrument shows adequate goodness of fit and retained stability and reliability, and is proposed as a valid and reliable means to evaluate and monitor critical thinking in university students.

Keywords: critical thinking; assessment; measurement; instruments; scale; validation studies; psychometrics; factor analysis; higher education students

Citation: Payan-Carreira, R.; Sacau-Fontenla, A.; Rebelo, H.; Sebastião, L.; Pnevmatikos, D. Development and Validation of a Critical Thinking Assessment-Scale Short Form. *Educ. Sci.* **2022**, *12*, 938. https://doi.org/10.3390/educsci12120938

Academic Editor: Sandra Raquel Gonçalves Fernandes

Received: 18 November 2022
Accepted: 15 December 2022
Published: 19 December 2022

1. Introduction

Improving critical thinking (CrT) skills remains a growing concern for today's Higher Education Institutions (HEI). CrT is a crucial non-technical, soft skill, highly prized by stakeholders in every profession, which led to a market-driven educational culture. CrT has been identified as one of the top soft skills sought in the twenty-first century [1–4]. The HEIs raised the purpose of nurturing their students in critical thought and informed decision-making to provide the market with a skilled workforce and thereby improve their employment rates [5].

CrT involves a complex combination of higher-order reasoning processes. More than the sum of individual skills, CrT is perceived as being interwoven of various multidimensional and multi-levelled sub-skills. For instance, within the Think4Jobs ERASMUS+ project, a working definition was conceptualised under the Facione framework [6] as the "*purposeful mental process driven by conscious, dynamic, self-directed, self-monitored, self-corrective thinking, sustained by disciplinary and procedural knowledge as well as metacognition*" [7].

CrT conceptualization has diverged through time, in accordance with three large branches: philosophical, psychological and educational [8]. For the philosophical approach,

focused on the mental process of thought, a critical thinker is someone that logically evaluates and questions the assumptions of others and his own, while for the psychological approach, focused on the processes driving an action, the critical thinker holds a combination of skills that allow individuals to assess a situation to decide on the best action to take. The educational approach places itself closer to the psychological approach, and relies on the use of frameworks and learning activities designed to enhance students' CrT skills, and consequently to test these skills [8,9].

CrT requires a complex set of qualities that may be foreseen as "generic" or as "domain-specific" [10,11]. The generic-CrT-skills usefulness transcends the academic and professional settings, and applies to all aspects of one's life; it is particularly foreseen to judge challenging moral and ethical situations that are often framed by particular interests [3,12]. Domain-specific CrT skills are often framed by a standard intervention or a code of professional conduct as expected from professionals, and support the decision-making within a particular context. Furthermore, most concepts recognize that CrT embeds in abilities or skills supported by a set of sub-skills, as well as in attitudes or dispositions [13–15]. The dispositions comprise different dimensions, and they determine whether a person is willing to use critical thinking in everyday life.

For HEI, a challenge exists regarding CrT development in their students: 1—how can they be efficiently and effectively taught along with the programme curricula, to mitigate putative gaps regarding the expectations of stakeholders; and 2—how can they be assessed to both validate the strategies' effectiveness and to demonstrate the students' acquisition of CrT skills? Educators in HEI have been confronted with the need to adopt appropriate teaching strategies to enhance students' CrT skills and assessment tools to show evidence of students' achievements in this regard [16]. Another challenge faced by HEI and educators is respecting the complexity of CrT' nature, which should be made explicit to students, while avoiding the pressure that may be associated with the reported limitations of the "teach-to-test" approach [17], improving the odds of developing and transferring CrT skills to everyday life or the labour market.

Regarding the evaluation of CrT skills, two different approaches have been used [18]. One approach uses resources such as different types of measuring instruments, either the formal or standardized CrT tests (such as the CCTT—Cornell Critical Thinking Test; the California Critical Thinking Dispositions Inventory—CCTDI; or the Halpern Critical Thinking Assessment test—HCTA, among others) [8], or the self-reported students' or stakeholders' perceptions [8]. The other approach uses "objective measurements" or "performance assessment", which are based on the transferability of skills to new work-based, professional-driven situations (e.g., the PBDS—Performance-Based Development System Test for nursing [19], the OSCE—Objective Structured Clinical Examination for clinical subjects [20], or the iPAL—Performance Assessment of Learning for engineering [11]). The performance assessment combines different dimensions of technical and soft-competencies evaluation. In general, performance-based critical-thinking tests rely on simulated real-life decision-making situations, demanding the students present the rationale for their decisions, using the available evidence [21].

Whether standardized or self-reported, CrT tests share a common pitfall: they tend to assume that critical thinking can be fragmented into a sum of detached sets of measurable sub-skills, such as analysis, evaluation, inference, deduction and induction [10]. According to those defending the performance assessment, there is little support that CrT sub-skills, or even the CrT skill for that effect, are independently mobilized in everyday life or work contexts. Therefore, performance assessment allows for a holistic evaluation of a set of CT skills or sub-skills combined differently, to succeed in the task in hand.

According to Simper et al. [22], "*Critical thinking, problem solving and communication are fundamental elements of undergraduate education, but methods for assessing these skills across an institution are susceptible to logistical, motivational and financial issues*". Standardized tests are based on well-established CrT-skills taxonomies, such as Ennis' and Facione's [8], and have been used for a long time to measure the CrT skills in students or HEI candidates worldwide.

Even though some of these tools were validated at the time, they have been recently questioned regarding the transferability of the construct validity across disciplines or regions [23,24], questioning their face validity. Moreover, they are not easily available; some demand expert evaluation and scoring, the rater needs to be trained [3], and are usually expensive to routinely apply [25]. In addition, for some of them, the situations around the questionnaire are far from the students' reality or take between 50 to 80 min to respond [23], contributing to the poor motivation of respondents to fill in the questionnaires [23,26]. Other concerns include the use of forced-choice questions, which may restrict the respondent's answers by limiting the possible hypothesis and relying on recognition memory [27,28]; the fact that the questions are often constructed from inauthentic situations [8], designed to trigger a response from the respondent; and the possible limited relevance of the skills tested, compared with the proposed instruction outcomes [29]. Finally, at least for some particular tests, it remains unclear how the respondent's reasoning will allow evidence of more discrete dispositions, such as open-mindedness or inquisitiveness [8], or how they could avoid the use of specific reasoning skills in students positioned in the more advanced years of their academic path. Therefore, their use in an academic context remains controversial. In particular fields, discipline-specific formal tests, such as in Business and in Health Science, have been developed, to copy with a less generalist scope the CrT-skills instruments, but their use also involves costs, and they need to be further validated in other regional or cultural contexts.

Consequently, the usefulness of such instruments has been questioned regarding their regular application in academic contexts, particularly in assisting students' improvement of CrT, as a whole, or as specific skills, and as evidence of student progression across the curricula [16,30]. On the other hand, a mismatch may arise from differences between the tasks students must develop during learning and what is assessed by the standardized tests, leading to a situation where the assessment does not cope with the subject outcome.

In the past decades, dissatisfaction with standardized tests led to the development of self-report instruments that have been validated in various disciplines. Nevertheless, there are some differences between these tools in the conceptualization of the CrT skills underlying the construct, so they are not entirely equivalent, and may even be scoring different sets of competences. Consequently, it is challenging to establish a comparison between them. Self-report tests for critical thinking seem more frequently used to ascertain respondents' dispositions, rather than skills.

However, students' self-report data might not be a consensual approach. Cole and Gonyea's work showed that the scores obtained by the same students in standard and self-reported questionnaires often present low correlations, as students tend to overscore their performance in the self-report questionnaires [31]. Such a bias may be associated with the willingness to cope with social standards or to meet the teacher's expectations, or influenced by the need to evoke past events to answer a question [32]. Nonetheless, if there is awareness of their nature and recognition of the underlying reasons for them to occur, these kind of biases may be prevented when designing a self-report instrument [32]. Self-reporting methods have been widely used to assess CrT skills gained after changes in instructional strategies [21,33]. However, both the complexity of the construct of critical thinking and the small population enrolled in the studies contribute to the poor reliability of the constructs, thereby reducing the validity of some tests. A similar concern applies to the performance tests. Nonetheless, they may assume particular interest in assessing non-cognitive characteristics [34] when there is no direct reflection on the students' grades or opportunities, namely in educational settings. In this context, self-report questionnaires may be used to monitor and enhance performance and to identify individual training needs [34].

The European Project "Think4Jobs" (2020-1-EL01-KA203-078797), currently ongoing, aims at expanding the collaboration between Labor Market Organizations (LMOs) and HEI to design, implement and assess the efficacy of CrT-blended-apprenticeships curricula developed for five disciplines (i.e., Veterinary Medicine, Teacher Education, Business

and Economics, Business Informatics, English as a Foreign Language). These curricula were designed to provide students with the opportunity to systematically train CrT skills and to stimulate their transfer into new situations arising from the labour market. This collaboration is foreseen as a flexible interface sustaining HEI and LMO collaboration, to provide a work-based context for developing graduates' CT (https://think4jobs.uowm.gr/, accessed on 9 November 2022). The changes in the CrT skills were tested in students participating in piloting courses using new CrT-embedding instructional strategies in different disciplines. The changes in the CrT skills were tested in students participating in piloting courses using new CrT-embedding instructional strategies in different disciplines such as Teacher Education (Greece), Business Informatics (Germany), English as a Foreign Language (Lithuania), Business and Economics (Romania) and Veterinary Medicine (Portugal). Based on previous experience and available literature, it was decided among partners to abandon classical, standardized CrT-skills tests, and instead select a test that may cope with some primary criteria: a closed-end test; easy to administer online; matching the proposed outcomes of the activities that would be implemented with students to reinforce CrT skills, and covering the CrT skills as conceptualized under the Facione framework [35]; practical for students to take; and not demanding, in terms of the level of technical expertise required to answer and to retrieve information. In addition, a limit expected time for completion of the questionnaire was set (preferably less than 30 min), as it was intended to be used paired with a different questionnaire tackling CrT dispositions.

Among the questionnaires addressing the core CrT skills as conceptualized by Facione [13,35], namely interpretation, analysis, evaluation, inference, explanation, and self-regulation, each one encompassing subskills, the consortium selected the questionnaire developed by Nair during her doctoral thesis [36]— the Critical Thinking Self-Assessment Scale (CTSAS)—to be applied pre- and post-test to the students enrolled in the activities. The instrument was one of the first validated scales for self-evaluation of CrT skills in higher-education students, and was designed to be applied across different disciplines. The original final version was composed of 115 items scored according to a seven-point rating scale (ranging from 0 = never to 6 = always). The questionnaire has been tested in different geographic and cultural contexts, and scored well in the reliability and internal consistency tests, as well as in the confirmatory factor analysis for all the skills composing the questionnaire [37].

However, even though the expected time to complete Nair's questionnaire was around 50 min, according to the author [36], the time for filling in the questionnaire was longer when it was tested with a small group of students, and was slightly longer than desired. Consequently, it was decided to shorten the original scale, to reach a response time of less than 30 min.

The purpose of this study is to present and validate the psychometric characteristics of a new, short form of Nair's CTSAS, intended to assess the CrT skills of students engaged in activities designed to support the enhancement of CrT skills, to diagnose the skills needing intervention and to monitor the progress or the results of interventions.

2. Materials and Methods

2.1. Shorthening of the CTSAS

To shorten the original Nair scale composed of 115 items, a two-step approach was used, involving two Portuguese experts. The following criteria were outlined for the possible rejection of items: 1—low loading-weights elimination (items with loading weights below 0.500 were eliminated, with 84 items remaining); 2—elimination of redundant items and items whose specific focus was not set on the use of cognitive skills (since the partnership considered that 84 items was still a high number, the items considered as redundant or not focusing on cognitive skills were marked for elimination, and items were reduced to 58); 3—review by two experts (after marking the items for elimination, the proposal was analysed by two independent experts who confirmed or reverted the rejection proposal, based on the Facione-based conceptualization of CrT skills and subskills.

As recommended by the experts, the final version also incorporated items 16 and 19 from the original scale, due to their theoretical relevance. Modification of the items of the original CTSAS was avoided. Table 1 summarizes the changes introduced in the original questionnaire.

The CTSAS short form retained a total of 60 peer-reviewed items. The number of items assessing each dimension ranged between 7 and 13. For subdimensions (or subskills), this number varied from 3 to 7 items, except for 5 subdimensions (decoding significance, detecting arguments, assessing claims, stating results, and justifying procedures), which comprised only two items. The short-form scale retained the original scale's framework, where students start with the question «*What do you do when presented with a problem?*» and are requested to answer the items using a seven-point Likert-scale structure with the following options: 0 = Never; 1 = Rarely; 2 = Occasionally; 3 = Usually; 4 = Often; 5 = Frequently; 6 = Always.

Table 1. Comparison between the original Nair's CTSAS questionnaire and its short form.

CTSAS Dimensions (Skills/Sub-Skills)		Items in the Original CTSAS	Eliminated Items	Items in the CTSAS Short-Form
Interpretation	Categorization	1–9	2, 4, 6–8	1–3
	Clarifying meaning	15–21	18–20	6–9
	Decoding significance	10–14	10, 12, 14	4, 5
Analysis	Detecting arguments	28–33	32, 33	15, 16
	Analyzing arguments	34–49	34, 39	17–20
	Examining ideas	22–27	27–29	10–14
Evaluation	Assessing claims	40–44	40–42	21, 22
	Assessing arguments	45–52	46, 50, 52	23–27
Inference	Drawing conclusions	67–74	67, 68, 73	36–40
	Conjecturing alternatives	60–66	62, 65	31–35
	Querying evidence	53–59	53, 54, 58, 59	28–30
Explanation	Stating results	75–79	76, 77, 79	41, 42
	Justifying procedures	80–88	81, 83–88	43, 44
	Presenting arguments	89–96	95, 96	45–50
Self-regulation	Self-examination	97–105	98, 104	51–57
	Self-correction	106–115	107, 109–111, 113–115	58–60

2.2. Participants

Five hundred and thirty-one university students (389 women, 142 men) participated in this study, ranging from 19 to 58 years old (mean = 23.47; SD = 7.184). The distribution of participants by country was as follows: 33.3% were from Greece, 29.4% from Portugal, 21.1% from Romania, 9.8% from Lithuania and 6.4% from Germany. Students studied within the following disciplines: Business Informatics, Business and Economics, Teacher Education, English as a Foreign Language, and Veterinary Medicine.

Ethical clearance for the study was obtained from the University of Évora Ethical Committee (registered with the internal code GD/39435/2020); moreover, students signed an informed consent associated with the questionnaire, and were allowed to withdraw from the study at any time without penalty or loss of benefits.

2.3. Instruments and Procedures

2.3.1. Translation of the CTSAS Short Form into Different Languages

The adopted short-version of the CTSAS in English, was translated into Portuguese, Romanian, Greek and German. The translation into these languages followed the recommended procedures (translation, revision and refinement), to ensure that the meaning, connotation and conceptualization respected the original instrument [38,39]. Two bilingual translators from each country using a non-English-version questionnaire, converted the

adopted CTSAS short form into their mother language; different sets of operators then analysed this translation to screen the differences between the two versions of the questionnaire and ensure the precision of the translation and its compliance with the original [40]. The consensual translated versions were reviewed by a group of experts from each national team in the project, who judged the content equivalence of the instrument. The experts' concordance was considered as an equivalent assessment of the translated questionnaire.

2.3.2. Data Collection

The collection of data through the CTSAS short form was performed from October 2021 to January 2022, in accordance with the scheduled term in the different piloting courses designed in the Think4Jobs project. This study used a non-randomised, non-probability convenience sample resulting from the voluntary responses from students enrolled on the Think4Jobs' designed curricula. The participants were students from Greece (enrolled on the courses Teaching of Science Education, Teaching of the Study of the Environment, and Teaching of Biological Concepts), students from Germany (enrolled on the courses Design Patterns, Innovation Management, Economic Aspects of Industrial Digitalization, and Scientific Seminar), from Lithuania (enrolled on the English for Academic Purposes course); from Portugal (enrolled on the courses Deontology, Gynaecology, Andrology and Obstetrics, Imaging, and on Curricular Traineeship), and students from Romania (enrolled on the courses Pedagogy and Didactics of Financial Accounting, Business Communication, and Virtual Learning Environments in Economics), all of whom responded to questionnaires in Greek, German, English, Portuguese and Romanian, respectively.

The questionnaire was made available to students online, on the Google Forms platform. The invitation to participate was sent to the students at the beginning of the semester, through the course page on Moodle. The process was supervised by the teachers involved in the pilot courses.

The responses were collected into an individual Excel file for each country: after data anonymization (by replacing the names with an alpha-numeric code (composed of the code for the country—GR, LT, RO, GE and PT, respectively, for Greece, Lithuania, Romania, Germany and Portugal—plus a sequential number, from 1 to n), the removal of all other identifying information retrieved from the platform, and screening for inconsistent data, the files were merged into the database used for statistical analysis.

2.4. Statistical Analysis

The descriptive measures for the items included the mean, standard deviation, skewness, kurtosis, the equal distribution Kolmogorov–Smirnov test and the Mann–Whitney U test for the mean rank differences.

To assess if the CTSAS short form fits the original factor model, a confirmatory factor analysis (CFA) was performed, with weighted least-square means and variances (WLSMV) as an estimation method, due to the ordinal nature of the data [41]. Model fit-indices performed include the χ^2 test for exact fit, the comparative fit index (CFI), the Tucker–Lewis index (TLI) and the root-mean-square error of approximation (RMSEA). Following Hu and Bentler [42], we considered CFI and TLI values ≥ 0.90 and RMSEA ≤ 0.06 (90%IC) as acceptable fit values. Data were specified as ordinal in the model.

To evaluate the reliability and internal consistency of the scale and subscales, Cronbach's alpha was computed. In accordance with Hair et al. [43], we considered alphas above 0.70 as good reliability-indices.

The multigroup invariance was assessed for female and male students. Differences on RMSEA and CFI values lower than 0.015 and 0.01, respectively, were used as criteria for invariance [44,45].

Univariate descriptive- and internal-consistency was calculated using the IBM SPSS Statistics 26. CFA and multigroup invariance analysis were performed, using MPlus 7.4 [46].

3. Results

The results are divided into three sections. The first section presents the descriptive statistics of the items. The second section shows the results from the confirmatory factor analysis. The third section shows the multigroup invariance analysis.

3.1. Descriptive Analysis of Items

The mean range of the 60 items varied from 3.13 («*I write essays with adequate arguments supported with reasons for a given policy or situation*») to 5.04 («*I try to figure out the content of the problem*»). The standard deviation varied from 0.958 («*I try to figure out the content of the problem*») to 1.734 («*I write essays with adequate arguments supported with reasons for a given policy or situation*»). The K–S test showed that data were equally distributed for female and male students ($p > 0.05$), except for the item «*I can logically present results to address a given problem*» (Z = 1.533; $p = 0.018$) and the item «*I respond to reasonable criticisms one might raise against one's viewpoints*» (Z = 1.772; $p = 0.004$). The item descriptions are displayed in Table A1 (see Appendix A, Table A1).

The Mann–Witney U test showed no statistically significant differences between female and male students ($p > 0.05$) except for the items «*I observe the facial expression people use in a given situation*» (Std U = −2.230; $p = 0.026$), «*I can logically present results to address a given problem*» (Std U = 2.382; $p = 0.017$), «*I respond to reasonable criticisms one might raise against one's viewpoints*» (Std U = 3.957; $p < 0.001$) and «*I provide reasons for rejecting another's claim*» (Std U = 2.588; $p = 0.010$). Detailed item descriptions can be consulted in Appendix A, Table A1.

3.2. Confirmatory Factor Analysis (CFA) and Reliability

The aim of the CFA was to confirm whether the CTSAS short form (60 items) fitted the original second-order factor model proposed by Nair [36]. Five successive models of increasing complexity were tested to achieve a comprehensive analysis of the structure and relations of the sixty items, six latent skills and a general construct. Five successive models of increasing complexity were tested to achieve a comprehensive analysis of the structure and relations of the sixty items, six latent skills and a general construct:

1. Model 1: One-factor model. This model tests the existence of one global factor on critical thinking skills, which explains the variances of the 60 variables.
2. Model 2: Six-factor (non-correlated) model. This model tests the existence of six non-correlated factors that explain the variance of the set of items.
3. Model 3: Six-factor (correlated) model. This model tests the existence of six correlated latent factors, each one explaining the variance of a set of items.
4. Model 4: Second-order factor model. This model represents the original model proposed by Nair [36], in which a global critical-thinking-skills construct explains the six latent-skills variance, which, in turn, each explain a set of items.
5. Model 5: Bi-factor model. This model tests the possibility that the 60 scale-items variances are being explained by a global critical-thinking-skills construct, and by the six latent skills, independently.

Table 2 shows model fit-indices for each model. Goodness-of-fit indices are satisfactory for models 3 and 4, but not for models 1, 2 and 5. As model 3 and model 4 are not nested, we guide our interpretation based on fit-indices differences. The differential values of the RMSEA and CFI indices between model 3 (which shows the best goodness-of-fit indices) and model 4 (which represent the original model proposed by Nair [36]) are lower than 0.015 and 0.010, respectively (ΔRMSEA = 0.002; ΔCFI = 0.003), suggesting that both models may be used to validate the internal structure of the questionnaire. As model 4 represents the original model, it will be accepted as a fitted factor structure, and considered for the following analysis.

Table 2. Goodness-of-fit indices.

Models	χ^2 (df)	*p*	RMSEA [90%IC]	CFI	TLI
Model 1: 1-factor model	5159.412 (1710)	<0.0001	0.061 [0.059–0.063]	0.893	0.890
Model 2: 6-factor model (non-correlated)	29275.338 (1710)	<0.0001	0.174 [0.172–0176]	0.148	0.118
Model 3: 6-factor model (correlated)	3871.243 (1695)	<0.0001	0.049 [0.047–0.051]	0.933	0.930
Model 4: second-order factor model	3975.885 (1704)	<0.0001	0.051 [0.049–0.053]	0.927	0.924
Model 5: bi-factor model	18,656.904 (1657)	<0.0001	0.139 [0.137–0.141]	0.474	0.439

Factor loadings presented in Table A2 (see Appendix A, Table A2) are significant ($p < 0.001$) and vary from 0.386 («*I observe the facial expression people use in a given situation*») to 0.786 («*I continually revise and rethink strategies to improve my thinking*»). All factor loadings are above 0.50, except for the items «*I observe the facial expression people use in a given situation*» (0.386), «*I clarify my thoughts by explaining to someone else*» (0.422) and «*I confidently reject an alternative solution when it lacks evidence*» (0.470).

The internal consistency of the CTSAS short from is excellent (Cronbach's $\alpha = 0.969$). As shown in Table A3 (see Appendix A, Table A3), Cronbach's alphas for each scale are above 0.70, showing good factorial reliability. Correlations between factors and between factors and the general critical-thinking-skills construct are strong (from 0.750 to 0.965) (Table 3). All correlations are significant at *p*-value < 0.0001.

Table 3. Cronbach's alfa reliability index and correlations between factors and between the factors and the general critical-thinking-skills construct.

Skills	α	CrT-Skills	1	2	3	4	5
1. Interpretation	0.772	0.881					
2. Analysis	0.888	0.925	0.905				
3. Evaluation	0.858	0.965	0.810	0.934			
4. Inference	0.905	0.956	0.806	0.858	0.937		
5. Explanation	0.853	0.907	0.765	0.825	0.864	0.868	
6. Self-regulation	0.905	0.851	0.750	0.750	0.781	0.841	0.805

3.3. Multigroup Invariance

A multigroup invariance analysis was produced to verify the factorial-structure invariance across sexes. Multigroup invariance was tested using the WLSMV as an estimation method, due to the ordinal nature of the data. As an initial step, the baseline was created for both groups (female and male students) using independent CFAs for each group. After the baseline was created, a CFA was applied to both groups simultaneously, to test for invariance. We tested the three invariance models, from the less restrictive (the configural model) to the most restrictive (the scalar invariance). The results are shown below, in Table 4.

Table 4. The goodness-of-fit indices for multigroup invariance, by sex.

Baseline Models	**χ^2 (df)**	***p***	**RMSEA [90%IC]**	**CFI**	**TLI**
Female	3488.157 (1704)	<0.0001	0.052 [0.049–0.054]	0.929	0.926
Male	2314.349 (1704)	<0.0001	0.050 [0.045–0.055]	0.948	0.946
Invariance	**χ^2 (df)**	***p***	**RMSEA [90%IC]**	**CFI**	**TLI**
Configural invariance	5521.460 (3390)	<0.0001	0.049 [0.046–0.051]	0.939	0.936
Metric invariance	5490.717 (3444)	<0.0001	0.047 [0.045–0.050]	0.941	0.940
Scalar invariance	5613.987 (3732)	<0.0001	0.044 [0.041–0.046]	0.946	0.949
Model comparison	**χ^2 (df)**	***p***	**ΔRMSEA**	**ΔCFI**	
Metric vs. Configural	45.988 (54)	0.773	0.002	0.002	
Scalar vs. Configural	370.658 (342)	0.137	0.005	0.007	
Scalar vs. Metric	328.786 (288)	0.049	0.003	0.005	

Based on the goodness-of-fit values of the different invariance models tested (configural, metric and scalar), the stability of the factor structure in both sexes is confirmed. The difference (Δ) in CFI and RMSEA values between the models is less than 0.015 and 0.010, respectively, revealing the invariance of the factorial structure, the invariance of factor loadings and the invariance of the item intercepts when comparing female and male students.

Once the measurement invariance was confirmed, we tested the structural invariance related to the populational heterogeneity, as well as the latent-mean invariance. Structural invariance tests whether the covariance level between factors is the same for both groups. Latent-mean invariance assesses whether the latent means are equal in both groups.

Table 5 displays the results from the structural invariance in both groups. The Wald test shows a significant difference between factor correlations of the female and male models (Wald = 6.507; df = 1; $p = 0.011$). As seen in Table 5, factor covariances are significantly higher for the male model than for the female model, showing some population heterogeneity [47].

Table 5. Factor covariances by sex.

Skills	**Interpretation**		**Analysis**		**Evaluation**		**Inference**		**Explanation**	
	F	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
Analysis	0.888	0.941								
Evaluation	0.760	0.900	0.922	0.955						
Inference	0.759	0.890	0.838	0.902	0.924	0.956				
Explanation	0.739	0.849	0.816	0.877	0.850	0.907	0.856	0.925		
Self-regulation	0.720	0.808	0.738	0.780	0.759	0.825	0.805	0.907	0.782	0.885

F = Female students, M = Male students. All correlations are significant at *p*-level < 0.001.

Within the means invariance analysis, female students are the baseline group, with a latent mean equal to zero. The mean comparison is presented in Table 6. There are non-significant differences in factor means between females and males.

Table 6. Latent-means differences between female and male.

Skills	**ΔMeans**	**SE**	**Est/SE**	***p***
Interpretation	−0.014	0.106	−0.129	0.897
Analysis	0.023	0.096	0.244	0.807
Evaluation	0.071	0.096	0.736	0.462
Inference	−0.051	0.099	−0.512	0.608
Explanation	0.177	0.097	1.832	0.067
Self-regulation	−0.005	0.098	−0.046	0.963

4. Discussion

This study attempted to validate a short form of the CTSAS questionnaire originally developed by Nair [36]. The number of items was reduced from 115 to 60, to reduce the time for completion of the questionnaire, with a greater focus on cognitive processes. Even though Nair refers to having participants that took between 35 and 45 min to complete the questionnaire [36], and recommends a time of 40 to 50 min [36], a previous test with random researchers took them more than 60 min to fill in the original CTSAS form. The adaptation of the short form eliminated 55 items, reducing the time for completion of the questionnaire to a maximum of 30 min, while maintaining the original skills- and sub-skills-dimensions. Thus, it was possible to keep the six core-skills structure (Interpretation, Analysis, Evaluation, Inference, Explanation and Self-regulation).

The shortened form was tested with 531 students from five HEI disciplines, in five European countries. Data were collected during the first and second terms of the academic year 2021/2022. Country representativeness was skewed, as the Lithuanian and German groups had a smaller number of participants. Nonetheless, the total number of respondents was adequate for performing a robust confirmatory-factor analysis [48].

On average, the age of participants was close to 23.5 years old, ranging from 19 to 58 years. Close to 87% of the participants were aged below 31 years. The age distribution reflects the reality of the HEIs in the five countries represented in this study, where most students enter HEIs at around 18 to 19 years of age. Older students are less commonly found, and often represent non-traditional students who work while enrolled in college or who seek graduation programmes to adjust their careers or to acquire new competencies supporting economic improvement.

The sex of the respondents was unevenly distributed, with the females reaching close to 75% of the total participants. In Europe, females represent the majority among tertiary students, particularly in longer programmes and in masters' and doctoral cycles, even though differences in this trend are recorded in some countries. In general, females predominate in Business and Law and Administration, as well as in Health Sciences, Arts and Humanities, Social Sciences and Education. In contrast, in Engineering and Information and Technologies, males predominate [49]. Among the population enrolled in the current study, a small number of respondents belong to the Information and Technologies disciplines (German students, who represented 6.4% of the sampled students). Due to their numbers, they were insufficient to balance the females predominance in the other disciplines.

The CTSAS short form was validated through confirmatory factor analysis, the evaluation of internal consistency or reliability, and by testing the multigroup invariance for male and female students.

In the confirmatory factor analysis used to test the questionnaire dimensionality and accuracy, two models showed equivalent satisfactory goodness-of-fit indices, namely the correlated six-factor model (Model 3) and the second-order factor model (Model 4). The chi-square/*df* ratio in the second-order factor model and the correlated six-factor model (2.33 and 2.28, respectively), confirmed the overall fitness of both models, while the RMSEA value, together with the TLI and CFI indices, supported the very good fit of both models [43,50]. Therefore, both models may be used as adequate models for depicting the structure of the CTSAS short-form questionnaire.

The confirmatory factor analysis established the validity and reliability of the correlated six-factor empirical model for the CTSAS short form: Interpretation (nine items), Analysis (eleven items), Evaluation (seven items), Inference (thirteen items), Explanation (ten items) and Self-regulation (ten items). The Cronbach alphas of the overall instrument and of the six scales were high ($\alpha = 0.969$ for the overall scale and between 0.750 and 0.965 for the six factors), supporting the independent use of each one of the six skills-scales [27], whenever different dimensional aspects of CrT skills need to be evaluated separately. Nonetheless the correctness of assuming that CrT may be decomposed into a set of discrete, measurable skills, has recently been questioned [8]. A number of voices defend the fact that CrT is usually practised as an integrated competence, and it is incongruent and

potentially detrimental to reduce CrT to a series of skills [8]. Considering that CrT results from complex, multifactorial, interwoven and multileveled processes of thought [15], the second-order factorial model might better reflect the multidimensionality of CrT. Note here that the model that tested the hypothesis that all the 60 items are explained by one factor (model 1) or by the bi-factorial model (model 5) did not have an adequate fit to the data. That is, we cannot refer to critical thinking without referring to the six skills that constitute the higher-order concept of critical thinking. It also deals with the fact that the exercise of CrT is shaped by values and one's background, which adds a complexity to the development of CrT competences. Consequently, the integrated score provided by the CTSAS short form adequately recognizes the complex and dynamic interplay of the six skills measured by the instrument, and support a holistic assessment of the students' CrT skills. The second-order factorial model, which was used to establish the comparison of results with the original CTSAS questionnaire by Nair [36], also showed that only four items had a factorial load below 0.500 (items # 4, 6, 8, and 39), suggesting that all other items presented convergent validity [43]. Despite this, it was decided to keep the four questions, considering that the substantive contents they dealt with were relevant for the intended purposes.

A limitation of this study might be that a self-report instrument is proposed to test students' CrT skills, with all the inherent biases that might encompass such questionnaires [31,51]. However, this limitation may be overcome by using the aggregate level to report data for individual CrT skills or by using the global CrT score.

The overall CTSAS short form showed strong internal consistency, with a Cronbach's alpha of 0.969, suggesting the scale retained stability and reliability despite the reduction in the number of items in the instrument. In addition, the individual dimensions of the skills assessed with the CTSAS short form presented acceptable-to-good Cronbach's alpha values [51–54] of between 0.772 (Interpretation) and 0.903 (Inference). These coefficients suggest that the constructs measure the intended dimensions. In addition, the correlations between the total score of the CTSAS short form and the individual dimensions tested confirm that skills may be measured through the items retained in the new, shortened CTSAS version.

Strong positive significant correlations were found between skills, and between the skills and the overall CTSAS short form. This finding supports the existence of a good item-related validity that strengthens the good internal-consistency-reliability that was found.

Sex did not affect most data distribution, except for four particular items (4, 42, 47 and 50). Moreover, the CTSAS short-form maintained its factorial structure invariance across sexes, supporting its reliability for both genders.

In summary, the current study presents and validates a short-form CTSAS questionnaire to assess CrT skills and some subskills to be applied in academic contexts, with the learning activities designed under the Facione framework. The short-form questionnaire presents a good construct validity with a good model-data-fit, and very good overall reliability and validity when applied to a multinational population enrolled on five very different higher education programmes. The strengths of the correlations between the skills and between each skill and the whole scale, confirm the good reliability of the instrument.

Consequently, the short-form of the CTSAS is a comprehensive CrT-assessment tool which has the potential to be used by HEIs to assess CrT skills.

Author Contributions: Conceptualization, R.P-C., L.S. and D.P.; methodology and formal analysis, R.P.-C. and A.S.-F.; investigation, R.P.-C., H.R. and L.S.; resources and data curation, R.P.-C.; writing—original draft preparation, R.P.-C., H.R. and A.S.-F.; project administration and funding acquisition, R.P.-C. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the "Critical Thinking for Successful Jobs—Think4Jobs" Project, with the reference number 2020-1-EL01-KA203078797, funded by the European Commission/EACEA, through the ERASMUS + Programme. The European Commission support for the production of this publication does not constitute an endorsement of the contents, which reflect the

views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of University of Évora (GD/39435/2020, approved at 5 January 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Acknowledgments: Authors were thankful to the partners of the Think4Jobs project in the collection of data with the new instrument. The effort was coordinated by the head of each academic team: Juho Mäkiö at the University of Applied Sciences Emden Leer (Germany); Roma Kriaučiūnienė at the Vilnius University (Lithuania); Daniela Dumitru at the Bucharest University of Economic Studies (Romania), but appreciation is further extended to all the professors who collaborated in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Items descriptive statistics.

	Mean	Sd.	Skew.	Kurt.	K-S Test	*p*
1. I try to figure out the content of the problem.	5.04	0.958	−0.744	−0.232	0.152	1.000
2. I classify data using a framework.	3.89	1.319	−0.452	−0.140	0.994	0.276
3. I break the complex ideas into manageable sub-ideas.	3.96	1.357	−0.467	−0.049	0.718	0.682
4. I observe the facial expression people use in a given situation.	4.63	1.380	−1.071	0.715	0.914	0.374
5. I examine the values rooted in the information presented.	4.12	1.284	−0.532	−0.172	0.754	0.620
6. I restate another person's statements to clarify the meaning.	3.63	1.515	−0.359	−0.545	0.762	0.607
7. I figure out an example which explains the concept/opinion.	4.53	1.097	−0.785	0.550	0.601	0.863
8. I clarify my thoughts by explaining to someone else.	4.29	1.348	−0.803	0.203	0.864	0.445
9. I seek clarification of the meanings of another's opinion or points of view.	4.23	1.185	−0.483	−0.196	0.718	0.682
10. I examine the similarities and differences among the opinions posed for a given problem.	4.23	1.166	−0.742	0.765	0.518	0.951
11. I examine the interrelationships among concepts or opinions posed.	3.84	1.222	−0.364	0.101	0.629	0.823
12. I look for supporting reasons when examining opinions.	4.44	1.174	−0.692	0.436	0.640	0.808
13. I look for relevant information to answer the question at issue.	4.62	1.147	−0.855	0.657	0.651	0.790
14. I examine the proposals for solving a given problem.	4.65	1.089	−0.626	−0.100	0.260	1.000
15. I ask questions in order to seek evidence to support or refute the author's claim.	4.09	1.341	−0.566	−0.084	1.041	0.229
16. I figure out if author's arguments include both for and against the claim.	3.97	1.316	−0.433	−0.229	1.044	0.226
17. I figure out unstated assumptions in one's reasoning for a claim.	3.63	1.289	−0.287	−0.190	0.723	0.673
18. I look for the overall structure of the argument.	3.99	1.332	−0.580	0.136	0.864	0.444

Table A1. *Cont.*

	Mean	Sd.	Skew.	Kurt.	K-S Test	*p*
19. I figure out the process of reasoning for an argument.	4.02	1.306	−0.578	0.253	0.381	0.999
20. I figure out the assumptions implicit in the author's reasoning.	3.73	1.275	−0.436	−0.032	0.828	0.500
21. I assess the contextual relevance of an opinion or claim posed.	4.00	1.192	−0.493	0.387	0.810	0.528
22. I seek the accuracy of the evidence supporting a given judgment.	4.18	1.283	−0.693	0.306	0.858	0.453
23. I assess the chances of success or failure in using a premise to conclude an argument.	4.08	1.344	−0.599	−0.007	1.120	0.163
24. I examine the logical strength of the underlying reason in an argument.	4.06	1.295	−0.464	−0.030	0.919	0.367
25. I search for new data to confirm or refute a given claim	4.15	1.288	−0.644	0.142	0.708	0.698
26. I search for additional information that might support or weaken an argument.	4.34	1.195	−0.520	−0.206	0.435	0.992
27. I examine the logical reasoning of an objection to a claim.	4.17	1.310	−0.552	0.025	0.883	0.417
28. I seek useful information to refute an argument when supported by unsure reasons.	4.37	1.186	−0.655	0.478	0.314	1.000
29. I collect evidence supporting the availability of information to back up opinions.	4.21	1.317	−0.771	0.585	0.794	0.554
30. I seek for evidence/information before accepting a solution.	4.49	1.241	−0.729	0.176	0.355	1.000
31. I figure out alternate hypotheses/questions, when I need to solve a problem.	4.21	1.311	−0.645	0.166	1.042	0.228
32. Given a problem to solve, I develop a set of options for solving the problem.	4.33	1.255	−0.685	0.234	0.683	0.739
33. I systematically analyse the problem using multiple sources of information to draw inferences.	4.11	1.381	−0.596	−0.103	0.325	1.000
34. I figure out the merits and demerits of a solution while prioritizing from alternatives for making decisions.	4.01	1.320	−0.455	−0.130	0.812	0.525
35. I identify the consequences of various options to solving a problem.	4.36	1.208	−0.558	−0.009	0.625	0.830
36. I arrive at conclusions that are supported with strong evidence.	4.30	1.164	−0.328	−0.484	0.490	0.970
37. I use both deductive and inductive reasoning to interpret information.	4.00	1.330	−0.419	−0.259	0.766	0.600
38. I analyse my thinking before jumping to conclusions.	4.39	1.335	−0.710	0.065	0.437	0.991
39. I confidently reject an alternative solution when it lacks evidence.	3.89	1.417	−0.312	−0.587	0.541	0.932
40. I figure out the pros and cons of a solution before accepting it.	4.64	1.175	−0.721	0.216	0.710	0.695
41. I can describe the results of a problem using inferential evidence.	3.78	1.206	−0.269	0.068	0.701	0.709
42. I can logically present results to address a given problem.	4.18	1.138	−0.425	0.111	1.533	0.018
43. I state my choice of using a particular method to solve the problem.	4.03	1.277	−0.530	0.164	0.305	1.000
44. I can explain a key concept to clarify my thinking.	4.10	1.246	−0.408	−0.141	0.585	0.883

Table A1. *Cont.*

	Mean	Sd.	Skew.	Kurt.	K-S Test	*p*
45. I write essays with adequate arguments supported with reasons for a given policy or situation.	3.13	1.734	−0.208	−0.966	0.833	0.492
46. I anticipate reasonable criticisms one might raise against one's viewpoints.	3.92	1.319	−0.438	−0.340	0.730	0.661
47. I respond to reasonable criticisms one might raise against one's viewpoints.	3.82	1.292	−0.456	−0.055	1.772	0.004
48. I clearly articulate evidence for my own viewpoints.	4.22	1.159	−0.353	−0.283	0.195	1.000
49. I present more evidence or counter evidence for another's points of view.	3.61	1.338	−0.258	−0.540	0.664	0.770
50. I provide reasons for rejecting another's claim.	4.04	1.400	−0.535	−0.309	1.255	0.086
51. I reflect on my opinions and reasons to ensure my premises are correct.	4.43	1.136	−0.442	−0.421	0.540	0.932
52. I review sources of information to ensure important information is not overlooked.	4.26	1.317	−0.628	−0.074	1.009	0.260
53. I examine and consider ideas and viewpoints even when others do not agree.	4.20	1.156	−0.380	−0.235	0.174	1.000
54. I examine my values, thoughts/beliefs based on reasons and evidence.	4.41	1.159	−0.455	−0.151	0.143	1.000
55. I continuously assess my targets and work towards achieving them.	4.46	1.182	−0.472	−0.367	0.354	1.000
56. I review my reasons and reasoning process in coming to a given conclusion.	4.18	1.187	−0.349	−0.236	0.415	0.995
57. I analyze areas of consistencies and inconsistencies in my thinking.	4.01	1.294	−0.448	−0.192	0.926	0.358
58. I willingly revise my work to correct my opinions and beliefs.	4.27	1.263	−0.457	−0.172	0.663	0.772
59. I continually revise and rethink strategies to improve my thinking.	4.34	1.280	−0.601	−0.073	0.683	0.739
60. I reflect on my thinking to improve the quality of my judgment.	4.53	1.187	−0.805	0.752	0.235	1.000

Table A2. Items' loadings.

Item	Interpretation	Analysis	Evaluation	Inference	Explanation	Self-Regulation
1. I try to figure out the content of the problem.	0.662					
2. I classify data using a framework.	0.661					
3. I break the complex ideas into manageable sub-ideas.	0.633					
4. I observe the facial expression people use in a given situation	0.386					
5. I examine the values rooted in the information presented.	0.654					
6. I restate another person's statements to clarify the meaning.	0.499					
7. I figure out an example which explains the concept/opinion.	0.594					
8. I clarify my thoughts by explaining to someone else.	0.422					
9. I seek clarification of the meanings of another's opinion or points of view.	0.536					

Table A2. *Cont.*

Item	Interpret-ation	Analysis	Evaluation	Inference	Explanation	Self-Regulation
10. I examine the similarities and differences among the opinions posed for a given problem.		0.614				
11. I examine the interrelationships among concepts or opinions posed.		0.734				
12. I look for supporting reasons when examining opinions.		0.671				
13. I look for relevant information to answer the question at issue.		0.650				
14. I examine the proposals for solving a given problem.		0.701				
15. I ask questions in order to seek evidence to support or refute the author's claim.		0.666				
16. I figure out if author's arguments include both for and against the claim.		0.670				
17. I figure out unstated assumptions in one's reasoning for a claim.		0.619				
18. I look for the overall structure of the argument.		0.707				
19. I figure out the process of reasoning for an argument.		0.772				
20. I figure out the assumptions implicit in the author's reasoning.		0.745				
21. I assess the contextual relevance of an opinion or claim posed.			0.723			
22. I seek the accuracy of the evidence supporting a given judgment.			0.735			
23. I assess the chances of success or failure in using a premise to conclude an argument.			0.702			
24. I examine the logical strength of the underlying reason in an argument.			0.725			
25. I search for new data to confirm or refute a given claim			0.674			
26. I search for additional information that might support or weaken an argument.			0.732			
27. I examine the logical reasoning of an objection to a claim.			0.761			
28. I seek useful information to refute an argument when supported by unsure reasons.				0.717		
29. I collect evidence supporting the availability of information to back up opinions.				0.740		
30. I seek for evidence/information before accepting a solution.				0.691		
31. I figure out alternate hypotheses/questions, when I need to solve a problem.				0.734		
32. Given a problem to solve, I develop a set of options for solving the problem.				0.710		
33. I systematically analyse the problem using multiple sources of information to draw inferences.				0.738		
34. I figure out the merits and demerits of a solution while prioritizing from alternatives for making decisions.				0.742		
35. I identify the consequences of various options to solving a problem.				0.704		

Table A2. *Cont.*

Item	Interpretation	Analysis	Evaluation	Inference	Explanation	Self-Regulation
36. I arrive at conclusions that are supported with strong evidence.				0.756		
37. I use both deductive and inductive reasoning to interpret information.				0.696		
38. I analyse my thinking before jumping to conclusions.				0.636		
39. I confidently reject an alternative solution when it lacks evidence.				0.470		
40. I figure out the pros and cons of a solution before accepting it.				0.656		
41. I can describe the results of a problem using inferential evidence.					0.745	
42. I can logically present results to address a given problem.					0.749	
43. I state my choice of using a particular method to solve the problem.					0.672	
44. I can explain a key concept to clarify my thinking.					0.740	
45. I write essays with adequate arguments supported with reasons for a given policy or situation.					0.511	
46. I anticipate reasonable criticisms one might raise against one's viewpoints					0.606	
47. I respond to reasonable criticisms one might raise against one's viewpoints.					0.650	
48. I clearly articulate evidence for my own viewpoints.					0.720	
49. I present more evidence or counter evidence for another's points of view.					0.573	
50. I provide reasons for rejecting another's claim.					0.536	
51. I reflect on my opinions and reasons to ensure my premises are correct.						0.719
52. I review sources of information to ensure important information is not overlooked.						0.785
53. I examine and consider ideas and viewpoints even when others do not agree.						0.705
54. I examine my values, thoughts/beliefs based on reasons and evidence.						0.756
55. I continuously assess my targets and work towards achieving them.						0.673
56. I review my reasons and reasoning process in coming to a given conclusion.						0.728
57. I analyze areas of consistencies and inconsistencies in my thinking.						0.737
58. I willingly revise my work to correct my opinions and beliefs.						0.750
59. I continually revise and rethink strategies to improve my thinking.						0.786
60. I reflect on my thinking to improve the quality of my judgment.						0.763

Table A3. Cronbach' alpha for the CTSAS short form skills and sub-skills.

Skills	Alpha's Cronbach	Sub-Skills	Std Alpha's Cronbach
Interpretation	0.772	Categorization	0.670
		Clarifying meaning	0.673
		Decoding significance	0.473
Analysis	0.888	Detecting arguments	0.632
		Analyzing arguments	0.812
		Examining ideas	0.799
Evaluation	0.858	Assessing claim	0.723
		Assessing arguments	0.821
Inference	0.905	Drawing conclusions	0.743
		Conjecturing alternatives	0.843
		Querying evidence	0.752
Explanation	0.853	Stating results	0.688
		Justifying procedures	0.681
		Presenting arguments	0.778
Self-regulation	0.905	Self-examining	0.860
		Self-correction	0.834

References

1. Dumitru, D.; Bigu, D.; Elen, J.; Jiang, L.; Railienè, A.; Penkauskienè, D.; Papathanasiou, I.V.; Tsaras, K.; Fradelos, E.C.; Ahern, A.; et al. *A European Collection of the Critical Thinking Skills and Dispositions Needed in Different Professional Fields for the 21st Century*; UTAD: Vila Real, Portugal, 2018.
2. Cruz, G.; Payan-Carreira, R.; Dominguez, C.; Silva, H.; Morais, F. What critical thinking skills and dispositions do new graduates need for professional life? Views from Portuguese employers in different fields. *High. Educ. Res. Dev.* **2021**, *40*, 721–737. [CrossRef]
3. Braun, H.I.; Shavelson, R.J.; Zlatkin-Troitschanskaia, O.; Borowiec, K. Performance Assessment of Critical Thinking: Conceptualization, Design, and Implementation. *Front. Educ.* **2020**, *5*, 156. [CrossRef]
4. Cinque, M.; Carretero, S.; Napierala, J. *Non-Cognitive Skills and Other Related Concepts: Towards a Better Understanding of Similarities and Differences*; Joint Research Centre, European Commission: Brussels, Belgium, 2021; 31p.
5. Pnevmatikos, D.; Christodoulou, P.; Georgiadou, T.; Lithoxoidou, A.; Dimitriadou, A.; Payan Carreira, R.; Simões, M.; Ferreira, D.; Rebelo, H.; Sebastião, L.; et al. *THINK4JOBS TRAINING: Critical Thinking Training Packages for Higher Education Instructors and Labour Market Tutors*; University of Western Macedonia: Kozani, Greece, 2021.
6. Facione, P. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction (The Delphi Report)*; California Academic Press: Millbrae, CA, USA; Newark, DE, USA, 1990; 112p.
7. Payan-Carreira, R.; Sebastião, L.; Cristóvão, A.; Rebelo, H. How to Enhance Students' Self-Regulation. In *The Psychology of Self-Regulation*; Dutton, J., Ed.; Psychology of Emotions, Motivations and Actions; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2022; p. 22. (in press)
8. Rear, D. One size fits all? The limitations of standardised assessment in critical thinking. *Assess. Eval. High. Educ.* **2019**, *44*, 664–675. [CrossRef]
9. Thaiposri, P.; Wannapiroon, P. Enhancing Students' Critical Thinking Skills through Teaching and Learning by Inquiry-based Learning Activities Using Social Network and Cloud Computing. *Procedia-Soc. Behav. Sci.* **2015**, *174*, 2137–2144. [CrossRef]
10. Lai, R.E. Critical Thinking: A Literature Review. *Pearson Res. Rep.* **2011**, *6*, 40–41.
11. Shavelson, R.J.; Zlatkin-Troitschanskaia, O.; Beck, K.; Schmidt, S.; Marino, J.P. Assessment of University Students' Critical Thinking: Next Generation Performance Assessment. *Int. J. Test.* **2019**, *19*, 337–362. [CrossRef]
12. Pnevmatikos, D.; Christodoulou, P.; Georgiadou, T. Promoting critical thinking in higher education through the values and knowledge education (VaKE) method. *Stud. High. Educ.* **2019**, *44*, 892–901. [CrossRef]
13. Facione, P.A. The Disposition Toward Critical Thinking: Its Character, Measurement, and Relationship to Critical Thinking Skill. *Informal Log.* **2000**, *20*, 61–84. [CrossRef]
14. Ennis, R.H. The Nature of Critical Thinking: Outlines of General Critical Thinking Dispositions and Abilities. 2013. Available online: https://education.illinois.edu/docs/default-source/faculty-documents/robert-ennis/thenatureofcriticalthinking_51711_000.pdf (accessed on 17 November 2022).

15. Halpern, D.F. Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring. *Am. Psychol.* **1998**, *53*, 449–455. [CrossRef]
16. Nair, G.G.; Stamler, L. A Conceptual Framework for Developing a Critical Thinking Self-Assessment Scale. *J. Nurs. Educ.* **2013**, *52*, 131–138. [CrossRef]
17. Rapps, A.M. Let the Seuss loose. In *Rutgers*; The State University of New Jersey: Camden, NJ, USA, 2017.
18. Tight, M. Twenty-first century skills: Meaning, usage and value. *Eur. J. High. Educ.* **2021**, *11*, 160–174. [CrossRef]
19. Ryan, C.; Tatum, K. Objective Measurement of Critical-Thinking Ability in Registered Nurse Applicants. *JONA J. Nurs. Adm.* **2012**, *42*, 89–94. [CrossRef] [PubMed]
20. Patrício, M.F.; Julião, M.; Fareleira, F.; Carneiro, A.V. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Med. Teach.* **2013**, *35*, 503–514. [CrossRef] [PubMed]
21. Hyytinen, H.; Ursin, J.; Silvennoinen, K.; Kleemola, K.; Toom, A. The dynamic relationship between response processes and self-regulation in critical thinking assessments. *Stud. Educ. Eval.* **2021**, *71*, 101090. [CrossRef]
22. Simper, N.; Frank, B.; Kaupp, J.; Mulligan, N.; Scott, J. Comparison of standardized assessment methods: Logistics, costs, incentives and use of data. *Assess. Eval. High. Educ.* **2019**, *44*, 821–834. [CrossRef]
23. Verburgh, A.; François, S.; Elen, J.; Janssen, R. The Assessment of Critical Thinking Critically Assessed in Higher Education: A Validation Study of the CCTT and the HCTA. *Educ. Res. Int.* **2013**, *2013*, 198920. [CrossRef]
24. Hart, C.; Da Costa, C.; D'Souza, D.; Kimpton, A.; Ljbusic, J. Exploring higher education students' critical thinking skills through content analysis. *Think. Ski. Creat.* **2021**, *41*, 100877. [CrossRef]
25. Williamson, D.M.; Xi, X.; Breyer, F.J. A Framework for Evaluation and Use of Automated Scoring. *Educ. Meas. Issues Pract.* **2012**, *31*, 2–13. [CrossRef]
26. Haromi, F.; Sadeghi, K.; Modirkhameneh, S.; Alavinia, P.; Khonbi, Z. Teaching through Appraisal: Developing Critical Reading in Iranian EFL Learners. *Proc. Int. Conf. Current Trends Elt* **2014**, *98*, 127–136. [CrossRef]
27. Ku, K.Y.L. Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Think. Ski. Creat.* **2009**, *4*, 70–76. [CrossRef]
28. de Bie, H.; Wilhelm, P.; van der Meij, H. The Halpern Critical Thinking Assessment: Toward a Dutch appraisal of critical thinking. *Think. Ski. Creat.* **2015**, *17*, 33–44. [CrossRef]
29. Liu, O.L.; Frankel, L.; Roohr, K.C. Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment. *ETS Res. Rep. Ser.* **2014**, *2014*, 1–23. [CrossRef]
30. Hatcher, D.L. Which test? Whose scores? Comparing standardized critical thinking tests. *New Dir. Inst. Res.* **2011**, *2011*, 29–39. [CrossRef]
31. Cole, J.S.; Gonyea, R.M. Accuracy of Self-reported SAT and ACT Test Scores: Implications for Research. *Res. High. Educ.* **2010**, *51*, 305–319. [CrossRef]
32. Althubaiti, A. Information bias in health research: Definition, pitfalls, and adjustment methods. *J. Multidiscip Healthc.* **2016**, *9*, 211–217. [CrossRef]
33. Payan-Carreira, R.; Cruz, G.; Papathanasiou, I.V.; Fradelos, E.; Jiang, L. The effectiveness of critical thinking instructional strategies in health professions education: A systematic review. *Stud. High. Educ.* **2019**, *44*, 829–843. [CrossRef]
34. Kreitchmann, R.S.; Abad, F.J.; Ponsoda, V.; Nieto, M.D.; Morillo, D. Controlling for Response Biases in Self-Report Scales: Forced-Choice vs. Psychometric Modeling of Likert Items. *Front. Psychol.* **2019**, *10*, 2309. [CrossRef]
35. Nair, G. *Preliminary Psychometric Characteristics of the Critical Thinking Self-Assessment Scale*; University of Saskatchewan: Saskatoon, SK, Canada, 2011.
36. Nair, G.G.; Hellsten, L.M.; Stamler, L.L. Accumulation of Content Validation Evidence for the Critical Thinking Self-Assessment Scale. *J. Nurs. Meas.* **2017**, *25*, 156–170. [CrossRef]
37. Gudmundsson, E. Guidelines for translating and adapting psychological instruments. *Nord. Psychol.* **2009**, *61*, 29–45. [CrossRef]
38. Tsang, S.; Royse, C.F.; Terkawi, A.S. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J. Anaesth.* **2017**, *11*, S80–S89. [CrossRef]
39. Gerdts-Andresen, T.; Hansen, M.T.; Grøndahl, V.A. Educational effectiveness: Validation of an instrument to measure students' critical thinking and disposition. *Int. J. Instr.* **2022**, *25*, 685–700. [CrossRef]
40. Flora, D.B.; Curran, P.J. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* **2004**, *9*, 466–491. [CrossRef] [PubMed]
41. Hu, L.t.; Bentler, P.M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. A Multidiscip. J.* **1999**, *6*, 1–55. [CrossRef]
42. Hair, J.F.; Page, M.; Brunsveld, N. *Essentials of Business Research Methods*, 4th ed.; Routledge: New York, NY, USA, 2019.
43. Cheung, G.W.; Rensvold, R.B. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Struct. Equ. Model. A Multidiscip. J.* **2002**, *9*, 233–255. [CrossRef]
44. Chen, F.F. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Struct. Equ. Model. A Multidiscip. J.* **2007**, *14*, 464–504. [CrossRef]
45. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*; Muthén & Muthén: Los Angeles, CA, USA, 2012.
46. Brown, T.A. *Confirmatory Factor Analysis for Applied Research*, 2nd ed.; Guiford Press: New York, NJ, USA, 2015; 462p.
47. MacCallum, R.C.; Widaman, K.F.; Zhang, S.; Hong, S. Sample size in factor analysis. *Psychol. Methods* **1999**, *4*, 84–99. [CrossRef]

48. Commission, E. *Tertiary Education Statistics*; Eurostat: Luxembourg, 2022.
49. Feinian, C.; Curran, P.J.; Bollen, K.A.; Kirby, J.; Paxton, P. An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociol. Methods Res.* **2008**, *36*, 462–494. [CrossRef]
50. Rosenman, R.; Tennekoon, V.; Hill, L.G. Measuring bias in self-reported data. *Int. J. Behav. Healthc. Res.* **2011**, *2*, 320–332. [CrossRef]
51. Taber, K.S. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Res. Sci. Educ.* **2018**, *48*, 1273–1296. [CrossRef]
52. Marôco, J. *Análise de Equações Estruturais—Fundamentos Teóricos, Software & Aplicações*, 2nd ed.; ReportNumber, Análise e Gestão de Informação, Ltd.: Pero Pinheiro, Portugal, 2014; p. 390.
53. Maroco, J. *Análise Estatística com o SPSS Statistics*, 7th ed.; ReportNumber-Análise e gestão de Informação, Ltd.: Pero Pinheiro, Portugal, 2018; 1013p.
54. Clark, L.A.; Watson, D. Constructing validity: New developments in creating objective measuring instruments. *Psychol. Assess.* **2019**, *31*, 1412–1427. [CrossRef]