IJPES
www.ijpes.com
International Journal of Psychology and Educational Studies
ISSN: 2148-9378

Developing the Critical Thinking Skill Test for High School Students: A Validity and Reliability Study

Ali ORHAN[1], Şule ÇEVİKER AY[2]

[1]*School of Foreign Languages, Zonguldak Bülent Ecevit University, Zonguldak, Turkey* *0000-0003-1234-3919*

[2]*Faculty of Education, Düzce University, Düzce, Turkey* *0000-0002-9505-5105*

ARTICLE INFO

Article History
Received 02.05.2021
Received in revised form 04.10.2021
Accepted 20.10.2021
Article Type:Research Article

ABSTRACT

This research aimed to develop "Critical Thinking Skill Test for High School Students" to measure critical thinking (CT) skills of high school students. For the CT test prepared based on the sub-skills of inference, evaluating arguments, deduction, recognizing assumptions and interpretation which are deemed to represent CT (Watson and Glaser, 1994), content validity and face validity were achieved with expert opinions and the table of specifications. Following the item difficulty and item discrimination analyses performed to test the construct validity, 34 items were omitted from the test which was finalized with 51 items. While mean item difficulty values of the sub-tests vary between 0.51 and 0.63, mean item discrimination values range from 0.35 to 0.49. The total test has a mean item difficulty value of 0.52 and a mean item discrimination value of 0.42. Decision-making skill was used to test the criterion-related validity of the test. KR20 reliability coefficients calculated for the sub-tests ranged from 0.62 to 0.75. A KR20 reliability coefficient of the total test was 0.87. Moreover, a correlation coefficient of 0.84 was calculated with the split-half method. To test time invariance of the test, the correlation values calculated between the results of the two applications which were performed three months apart ranged from 0.57 to 0.70. The correlation coefficient for the total test is 0.70. Based on the results of validity and reliability studies, it can be said that the CT test will yield valid and reliable results in measuring CT skills of high school students.

Keywords:
Critical thinking skills, critical thinking test, measuring critical thinking skills, high school students

1. Introduction

Nowadays, access to information is quite easy. However, the absence of filter techniques for accuracy and precision of information shared is very apparent, and everyone can freely share information. Distinguishing the validity, accuracy, and reliability of this pile of data remains a challenge. Also, this makes it even harder for individuals to make right decisions. Particularly with the increasing use of media, the resistance to fake news is mitigated as well. According to the media literacy index data by Open Society Institute (2021), Turkey is the third country that is the least resistant to false news. Reportedly, such news is spread most easily in this country. Given that 75.3% of the Turkish population uses the internet and spend an average of 7 hours and 29 minutes daily on the internet (We Are Social, 2020), the magnitude of problems against the information explosion surely arises. Media is considered to be a centerpiece of the people's lives (Pérez Tornero & Varis, 2010) and a great part of their lives (Masterman, 1985). Hence, media inevitably manipulates them through either direct or indirect messages. An individual will either accept the given information without questioning or decide to accept or deny the information after examining its accuracy. Critical thinking (CT) skill is the greatest helper of accuracy examination. Indeed, in an era in which the

* This article was produced from Ali ORHAN's Phd thesis which was conducted under the supervision of Şule Çeviker Ay.

[1]**Corresponding author:** School of Foreign Languages, Zonguldak Bülent Ecevit University, Zonguldak, Turkey
e-mail: ali_orh_an@hotmail.com

Citation: Orhan, A. & Çeviker-Ay, Ş. (2022). Developing the critical thinking skill test for high school students: a validity and reliability study. *International Journal of Psychology and Educational Studies, 9*(1), 132-44. https://dx.doi.org/10.52380/ijpes.2022.9.1.561

influence of easily accessible information or several people toward one's thoughts is manifested, CT is individual's defense mechanism against the world (Epstein & Kernberger, 2012). On the other hand, the 21st-century skills introduced with the need for transformation that has been brought by the Fourth Industrial Revolution (Industry 4.0) have changed learner's roles and required individuals to possess high-order thinking skills, think effectively, and consequently be able to adapt to developments and innovations of the era. CT skill may be one of the most important skill among all high-order thinking skills (Presseisen, 1985). In fact, it is a functional way of thinking that involves reflective, reasonable, discreet, and logical decisions and is resorted to by individuals for making decisions and resolving future problems (Ennis, 1985; Norris & Ennis, 1989). Thus, through CT, individuals obtain sound and accurate information on what is going on around them. Moreover, they question, examine, and evaluate the collected information from their surroundings. This effective task of evaluation is about examining the underlying reasons and searching for solid evidence to attain the accuracy of information (Mason, 2008). In like manner, individuals assess the sensibility, truth, and accuracy of given information, claims, evidence, and judgments and draw a conclusion through CT (Lewis & Smith, 1993). Undoubtedly, CT is the act of challenging a piece of information received from others (Judge, Jones, & McCreery, 2009) and distinguishing between right and wrong by reasoning (Wood, 1998).

One should not only think of one skill when it comes to CT because it is a collective skill composed of several sub-skills or sub-dimensions (Fisher, 2001). Many classifications can be observed in the literature for the sub-skills of CT (Ennis &Weir, 1985; Chance, 1986; Paul, Binker, Jensen &Kreklau, 1990; Facione, 1990, 2000; Kennedy, Fisher &Ennis, 1991; Pascarella &Terenzini, 1991; Watson &Glaser, 1994; Swartz &Parks, 1994; Jones et al., 1995; Jonassen, 2000; Halpern, 2003; Ennis, Millman &Tomko, 2005). Although they may be given different names, interpretation, analysis, evaluation, inference and explanation seem to be included in all of those classifications. Differently, some of the classifications include skills such as self-regulation (Facione, 1990, 2000), reflection (Jones et al., 1995), and deduction (Watson &Glaser, 1994). As one thing that can be considered a shortcoming, some of these classifications include certain skills in a narrower fashion (e.g. Swartz and Parks (1994) limits the aspect of evaluation to the evaluation of sources only). Taken together, it can be said that Facione (1990, 2000), Ennis, Millman and Tomko (2005), and Watson and Glaser (1994) made the most inclusive classifications. While Facione (1990, 2000) address CT skills in six aspects of interpretation, analysis, evaluation, inference, explanation, and self-regulation, Ennis, Millman and Tomko (2005) divides CT skills into six aspects of induction, deduction, observation, semantics, assumption, and questioning the credibility of sources. According to Watson and Glaser (1994), CT includes the skills of inference, recognizing assumptions, deduction, interpretation, and evaluating arguments.

There are several instruments developed to measure CT skills or dispositions of different age groups in the literature. Some of them measure CT skills while others measure CT dispositions. One of the most common measures used for CT skills is Watson-Glaser Critical Thinking Appraisal (WGCTA). Developed by Watson and Glaser in 1964, number of test questions had been gradually reduced and its different forms had been published until 1994. Whereas its original form included 100 questions in 1964, the number of questions was downed to 80 in its 1980 forms and 40 in its 1994 forms (Watson & Glaser, 1964; Watson &Glaser, 1994). Turkish adaptation studies of the test forms were conducted by different researchers with high school students (Çıkrıkçı, 1993, 1996; Evcen, 2002) and undergraduates (Aybek &Çelik, 2007). After the adaptation study by Çıkrıkçı (1996), KR20 internal consistency coefficient for the total test was found to be 0.63 and ranged from 0.20 to 0.47 for sub-tests. In an another adaptation study by Evcen (2002), KR20 internal consistency coefficient for total WGCTA was found to be 0.46 and KR20 internal consistency coefficients for sub-tests ranged from 0.29 to 0.53. Developed by Ennis, Millman and Tomko (2005) in 1985, Cornell Critical Thinking Test has two forms which are Level X and Level Z. Level X was developed for younger students while Level Z was developed for students studying at high schools and universities. Turkish adaptation for level Z of Cornell Critical Thinking Test was conducted by Şenturan (2006), and its level X was adapted to Turkish language by Kurnaz (2007). While Şenturan (2006) calculated the KR20 internal consistency coefficient as 0.45, Kurnaz (2007) calculated as 0.58 for the total testafter their adaptation studies.Ennis-Weir Critical Thinking Essay Test that was developed for the undergraduate students was adapted to Turkish language by Koç (2007). Adapted to Turkish language by Mecit (2006), Cornell Conditional Reasoning Test was developed by Ennis and Millman to determine CT skills of elementary and high school students. Mecit (2006) calculated the internal consistency of the test as 0.75.Tests developed by Facione (1990) and Shipman

(1983) are also used to measure CT skills of high school students and undergraduates. As well as international tests for measuring CT skills, there are CT tests developed by Turkish researchers in the literature. For example, Eğmir and Ocak (2016) developed a CT test to measure CT skills of the fifth-graders. Similarly, Demir (2006) developed an instrument titled Critical Thinking Scales to measure CT skills of the fourth- and fifth-graders. Demir's (2006) scales aim to measure analysis, evaluation, inference, interpretation, explanation, and self-regulation sub-skills.

Considering the CT studies performed in Turkey overall, almost all of the studies seem to have benefited from international instruments that measure CT skills. Although those international instruments have been adapted to Turkish language by different people at different times, none of the Turkish studies utilized a Turkish culture-specific CT test. Indeed, while Turkish adaptation studies of these instruments developed on the basis of foreign cultures have been carried out, they are observed to have a construct that is incompatible with the Turkish culture. Understandably, their reliabilities are at the lowest possible acceptable levels particularly in some of the dimensions, and our culture is unfamiliar with the examples, names, and cases used in the tests. Consequently, this has an impact of a valid and reliable measurement of students' CT skills. Several studies in the literature suggest that a Turkish culture-specific CT instrument should be developed (Aybek, 2006; Gülveren, 2007; Kurnaz, 2007; Ay & Akgöl, 2008; Tufan, 2008; Yıldırım, 2010). Although there are instruments that measure CT skills of elementary school students (Demir, 2006; Eğmir and Ocak, 2016), there is no comprehensive instrument for the high school level. Therefore, this study aimed to develop a Turkish culture-specific CT test to measure CT skills of high school students.

2. Methodology

2.1. Research Model

This is a test-development study for a measurement tool to determine the high school students' CT skills.

2.2. Study Group

Tabachnick and Fidell (2012) argued that a pilot study needs to be carried out with at least 150 participants to conduct validity and reliability studies of a measure. Özçelik (2013) suggests that a draft test should be given to approximately 400 individuals, whereas Kline (2010) articulates that about 200 individuals suffice. In the literature, some of the researchers also argue that a draft measure should be applied to a sample group of a size five times (Stevens, 2009; Floyd & Widaman, 1995) or 10 times the number of items (Gorsuch, 2014). Thus, approximately 800 individuals were deemed sufficient to conduct validity and reliability studies for the 87-item draft CT test. Because some of the test forms might not be answered intently or answered incompletely as anticipated, the pilot study was conducted with about 1000 individuals as planned.

The pilot application of the draft test was conducted at the high schools in the city center of a province located in the West Black Sea Region. Pilot application of a measure requires a sample that can appropriately represent test's target population (Crocker & Algina, 1986). Multilevel cluster sampling method was utilized to select the sample group for the pilot application of the CT test to be used with high school students. Accordingly, different types of high schools constitute the first-level clusters, and the grade levels at those high schools constitute the second-level clusters. The sampling process is provided in Table 1.

Table 1. *Sampling for the Pilot Application*

First-level clusters (choosing the high schools that have the highest representation rate relative to the number of students)	FIRST-LEVEL CLUSTERS		Second-level clusters (choosing one 9th-, 10th-, 11th-, and 12th-grade classes randomly in each high school)	SECOND-LEVEL CLUSTERS	
	N	%			
Anatolian High School	603	%6.46		9-grade:	232
Anatolian High School	629	%6.74		10-grade:	245
Vocational High School	510	%5.47		11-grade:	239
Vocational High School	997	%10.69		12-grade:	229
Science High School	261	%2.80			
Religious Vocational High School	826	%8.85		TOTAL: 945	
Anatolian High School	529	%5.67			
Religious Vocational High School	571	%6.12			
TOTAL	4926	%52.80			

The first-level clusters of the study are composed of two Vocational High Schools, two Anatolian High Schools, two Religious Vocational High Schools, and one Science High School, and one Anatolian High School, which admit students by examination, in the city center of a province in the West Black Sea Region. These schools have the highest representation rate relative to the number of students. The total number of students in these selected high schools corresponds to 52.80% of the number of students in all high schools in the city center of the province. Thus, the first-level clusters represent all high school types. For the second-level clusters, each of 9th-, 10th-, 11th-, and 12th-grade classes in these high schools was assigned as sub-clusters, which was participated by 945 students. Of these students, 232 are 9th graders, 245 are 10th graders, 239 are 11th graders, and 229 are 12th graders. Test forms found to be completed carelessly and to involve a great number of incomplete data were not included in the validity and reliability studies. Of the students who completed 705 test forms that were included in the validity and reliability studies, 52.5% are female, and 47.5% are male. Moreover, 24.4% are 9th graders, 26.2% are 10th graders, 25.4% are 11th graders, and 24% are 12th graders. In addition, most of the students are mostly 16 (27%), 17 (25.1%), 15 (24.1%), and 14 years old (19.6%), respectively. Their distribution by high school types is as follows: Anatolian High School (56.2%), Science High School (15.6%), Religious Vocational High School (14.5%), and Vocational High School (13.8%).

2.3. Procudere

The following steps were followed to develop the CT test in compliance with the literature (Crocker & Algina, 1986; Cronbach, 1984):

1. determining the purpose of the test, psychological attribute which the test aims to measure, and the behaviors which represent that attribute
2. creating a table of specifications that shows the item ratios for the behaviors determined
3. creating an item pool
4. preparing the test form
5. receiving expert opinion, and performing the preliminary application of the test
6. performing the pilot application of the test
7. conducting the validity and reliability analyses
8. preparing the guidelines for application, assessment, scoring of the test, and interpretation of the scores

2.3.1.Determining the purpose of the test, psychological attribute which the test aims to measure, and the behaviors which represent that attribute

The CT test to be developed aims to measure CT skills of high school students. Scores to be obtained in the test will be used to determine students' CT skill levels. Thus, a thorough literature review was performed to determine the CT sub-skills and the behaviors that represent those sub-skills. There are several sub-skill classifications which are deemed to represent CT skills (Watson & Glaser, 1994; Facione, 1990, 2000; Jones et al., 1995; Kennedy, Fisher, & Ennis, 1991; Paul, Binker, Jensen, & Kreklau, 1990; Swartz & Parks, 1994; Pascarella & Terenzini, 1991; Ennis & Weir, 1985; Ennis, Millman, & Tomko, 2005; Chance, 1986; Halpern, 2003; Jonassen, 2000). Considering these classifications made by different researchers, the most inclusive classifications seem to have been made by Facione (1990, 2000), Ennis, Millman, and Tomko (2005), and Watson and Glaser (1994). Therefore, the classification by Watson and Glaser (1994) that is thought to be very inclusive and has been commonly recognized was utilized in this study. Accordingly, the sub-skills of inference, evaluating the arguments, deduction, recognizing the assumptions, and interpretation (Watson & Glaser, 1994), which are deemed to represent the CT skill, were bases in developing the CT test. Watson and Glaser (1994) define the inference sub-skill as inferring new information based on a certain piece of knowledge or a situation or drawing conclusions from a proposition that is accepted to be true. This task of inference occurs through deduction, induction, and reasoning. The sub-skill of recognizing assumptions refers to identifying the assumptions, familiarizing with them, and deciding whether it is possible to make that assumption based on the current situation. The deduction sub-skill is about new propositions logically and obligatorily drawn from propositions that are known or assumed to be true. The interpretation sub-skill refers to evaluating the evidence for a situation or the solution of a problem, drawing conclusions based on this evidence, and assessing the accuracy of these conclusions. The sub-skill of evaluating the arguments is about determining strengths or weaknesses of inferences, statements, judgments, and evidence.

2.3.2. Creating a table of specifications that shows the item ratios for the behaviors determined

Achieving the content validity of a test requires explicitly determining all the behaviors that represent the attribute to be measured and then writing down the items that can measure those behaviors (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014). The most reasonable way of providing the content validity of achievement tests is to create a table of specifications and receive expert opinion (Terzi, 2019; Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014). Therefore, a table of specifications was prepared in regard to the number of items with which each sub-skill would be measured in the CT test.

2.3.3. Creating an Item Pool

Prior to creating an item pool for the CT test, CT instruments published both abroad and in Turkey were reviewed in detail (Watson & Glaser, 1964; Eğmir & Ocak, 2017; Demir, 2006; Ennis & Weir, 1985; Ennis, Millman, & Tomko, 2005). An item pool of 169 multiple-choice questions covering the five sub-skills was created. Of the questions in the item pool, 45 aim to measure the inference sub-skill, 26 aim to measure the sub-skill of evaluating the arguments, 17 aim to measure the deduction sub-skill, 55 aim to measure the sub-skill of recognizing the assumptions, and 26 aim to measure the interpretation sub-skill.

2.3.4. Preparing the test form

The questions in the item pool were adapted to the test format, and a draft CT test was prepared. Özçelik (2013) suggested that the test form should include an instruction at the top about what is expected from the students, how to answer the questions, and what to consider when answering them. In case students have no idea how to answer the test questions, a few example questions and answers should be included at the beginning of the test. Accordingly, in the draft CT test composed of five parts, explanations were added to the beginning of each part about what is expected from students and how to answer the questions in that part. One example question and its solution were also added for each part. With a brief explanation and narration, it was explained at the beginning of the test that the questions were to be answered based on the anecdotes noted down by a high school student in her imaginary diary. Because the test was developed for high school students, by this means, it was ensured that the test would attract their attention.

2.3.5. Receiving expert opinion, and performing the preliminary application of the test

The researcher consulted experts for opinion to achieve the content and face validity of the draft CT test. Kline (2010) suggested that content validity is best tested by receiving expert opinion rather than performing statistical analyses. Therefore, the draft CT test was submitted to six faculty members and two Turkish teachers. The faculty members work in the fields of Curriculum and Instruction (4), Mathematics (1), and Assessment and Evaluation (1). Two of the faculty members working in the field of Curriculum and Instruction have carried out many studies on CT before. Upon receiving the experts' feedbacks, eight questions were difficult to understand and had two corrected answers, and five questions that were considered unsuitable for high school students' levels and inadequate were omitted from the test. Moreover, the number of questions with the same purposes was reduced based on the mutual feedbacks from the experts in regard to the redundancy of items. Consequently, the draft CT test was finalized with 87 questions that were considered more successful and inclusive. Consisting of 10 items in the inference sub-test, 12 items in the deduction sub-test, 16 items in the sub-test of evaluating arguments, 23 items in the sub-test of recognizing assumptions, and 10 items in the interpretation sub-test (87 items in total), the draft CT test was tested on a small group prior to the pilot application.

In the preliminary application performed with 30 ninth graders, the students were encouraged to ask anything about any part of the test that was ambiguous. Moreover, the duration they needed to complete the test was observed, and the students were asked for their views on the difficulty level of the test questions. Following the preliminary application, two statements in the explanation of inference and deduction sub-tests were clarified upon the feedbacks from the students. The students reported that the difficulty level of the test questions was suitable, and it was found that 70 min were sufficient to complete the test.

2.3.6. Performing the pilot application of the test

The pilot application process performed with a total of 945 students from the selected eight high schools with multilevel clustering method took 2 weeks to complete. An optical form was designed for the test, and

the students were asked to mark their answers on the optical form. This made sure the prevention of possible problems in the digitization of the data following the pilot application. Upon the ethical approval by Düzce University Ethical Committee of Human Research No. 2019/86 dated November 5, 2019 and the research approval by Directorate of National Education No. E23489630 dated November 27, 2019, the test was applied at the high schools and the classrooms on site. The students were informed of the research and told that the research would be conducted on a voluntary basis. They were given 70 minutes to complete the test. During the pilot application process, the students were supervised, and optical forms of the students who were observed to answer the test carelessly were marked.

2.3.7. Conducting the validity and reliability analyses

Regarding the extent to which a test can measure a target attribute without involving other attributes (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014), validity is addressed in four groups, namely content validity, face validity, criterion-related validity, and construct validity (Crocker & Algina, 1986; Terzi, 2019). Content validity and face validity of the CT test were achieved through expert opinions and the table of specifications prepared beforehand. In addition, Büyüköztürk, Çakmak, Akgün, Karadeniz, and Demirel (2014) argue that another way of testing the content validity is to examine the correlation between the results of the test to be developed and another test known to measure the same attribute and content. Therefore, WGCTA, a commonly used test, and the draft CT test were applied to 100 students 2 weeks apart. Of these tests, 16 forms were not included in the analysis because of incomplete or neglectful answers, and results of the remaining 84 tests were compared with Spearman's rank correlation coefficient because of non-normal distribution of the data ($p< 0.05$).

In criterion-related validity, the correlation between the scores obtained from the draft test and the results of a valid and reliable instrument that measures another attribute that is thought to be related to the measured attribute is calculated (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014). Thus, the decision-making skill that is associated with CT skill in the literature (Halpern, 2003; Moore, 2010; Norris & Ennis, 1989; Bailin, 1998) was used to test the criterion-related validity of the CT test. Accordingly, the Adolescent Decision-Making Questionnaire developed by Çolakkadıoğlu (2012) and the draft CT test were given to 137 students 2 weeks apart. The reliability coefficients calculated by Çolakkadıoğlu (2012) for the sub-scales of Adolescent Decision-Making Questionnaire ranged from 0.76 to 0.85. For this study, the reliability coefficients of sub-scales ranged from 0.68 to 0.83. It was assumed that students with high scores of vigilance as a decision-making style in the Adolescent Decision-Making Questionnaire would have high CT skills. Moreover, it was assumed that students with higher scores of complacency, panic, and cop-out dimensions as decision-making styles would have lower CT skills.

Construct validity of the CT test was tested with item analysis. The two most common statistics in the item analysis of a test are item difficulty and item discrimination (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014). Tekin (2019) argued that a test should involve items at different difficulty levels and mean difficulty of a test should be at a level of 0.50. In other words, the test should have an average difficulty level and include questions at varying difficulties from easier to harder levels (Kan, 2011). Similarly, Özçelik (2013) stated that mean difficulty of tests should be at a level of 0.55. Although the behavior and attribute to be measured by the item is a continuous variable by nature, once it is made into a discrete variable artificially as 1–0, it is more appropriate to use the point biserial correlation coefficient for the item discrimination value (Çokluk, Şekercioğlu, & Büyüköztürk, 2012; Kan, 2011). Therefore, the point biserial correlation coefficient was utilized to calculate item discrimination values of the test items. An item difficulty value between 0 and 0.39 refers to a difficult item, a value between 0.40 and 0.59 refers to a moderate item, and a value of 0.60 and above refers to an easy item (Özçelik, 2013). The possible lowest item discrimination value for the items of the draft CT test was determined to be 0.30, and items below that value were not included in the test.Other than item analysis, the correlation between the draft test and another instrument known to measure a similar attribute can be used to test the construct validity (Terzi, 2019). Hence, results of the WGCTAwere also utilized for the construct validity.

Defined as test scores' level of being free from random errors, reliability (Turgut, 1995) can be calculated with methods such as KR20, KR21, parallel forms, test–retest, and split-half methods (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014; Terzi, 2019). Reliability coefficients for the sub-tests and the

total test were calculated with KR20 equation. Furthermore, the CT test was administered to a group of 59 individuals 3 months apart to check its time invariance. Both applications were performed in the second class hour of the day, and all students were given the same duration to complete the test. Because the data obtained from both applications were not normally distributed ($p< 0.05$), the results were compared with Spearman's rank correlation coefficient.

2.4. Ethical

In this study, all rules stated to be followed within the scope of "Higher Education Institutions Scientific Research and Publication Ethics Directive" were followed.

Ethical Review Board Name: Düzce University Ethics Committee

Date of Ethics Evaluation Decision: 05.11.2019 Ethics Assessment Document Issue Number: 2019/86

3. Findings

3.1. Findings on the Validity Study

After the CT test's content and face validity had been achieved through expert opinions and the table of specifications, its construct validity was tested with the item analysis. Following the item analysis, 34 items whose item discrimination values were below 0.30 (Item 2, Item 3, Item 5, Item 6, Item 8, Item 10, Item 12, Item 13, Item 14, Item 17, Item 19, Item 21, Item 23, Item 27, Item 28, Item 34, Item 35, Item 38, Item 39, Item 41, Item 57, Item 61, Item 62, Item 63, Item 64, Item 65, Item 66, Item 69, Item 70, Item 72, Item 74, Item 77, Item 79, Item 82) were omitted from the test. Although two items (Item 36, Item 37) had acceptable levels of item difficulty and item discrimination, these questions were omitted from the test along with the texts because there remained no other questions about the relevant text. Table 2 presents the item difficulty and item discrimination values of the items that were kept in the test following the item analysis.

Table 2. *Item Difficulty and Item Discrimination Values of Items Kept in the Test Following the Item Analysis*

Item No	Item Difficulty	Item Discrimination	Item No	Item Difficulty	Item Discrimination
1	0.69	0.33	49	0.57	0.61
4	0.69	0.33	50	0.53	0.56
7	0.43	0.44	51	0.36	0.43
9	0.52	0.36	52	0.57	0.34
11	0.46	0.41	53	0.55	0.43
15	0.59	0.35	54	0.53	0.40
16	0.30	0.32	55	0.45	0.30
18	0.35	0.37	56	0.48	0.42
20	0.62	0.30	58	0.51	0.51
22	0.45	0.30	59	0.45	0.48
24	0.65	0.55	60	0.44	0.50
25	0.68	0.52	67	0.33	0.43
26	0.74	0.30	68	0.37	0.35
29	0.65	0.31	71	0.43	0.49
30	0.57	0.61	73	0.30	0.30
31	0.65	0.30	75	0.75	0.33
32	0.46	0.43	76	0.52	0.32
33	0.71	0.30	78	0.39	0.38
40	0.34	0.33	80	0.52	0.41
42	0.55	0.55	81	0.60	0.46
43	0.55	0.55	83	0.43	0.40
44	0.58	0.52	84	0.65	0.56
45	0.60	0.50	85	0.51	0.47
46	0.33	0.35	86	0.54	0.46
47	0.44	0.50	87	0.59	0.52
48	0.58	0.50	TOTAL	0.52	0.42

Table 2 shows the 51 items kept in the test with item difficulty values between 0.75 and 0.30. Arguably, the test involves items at three different difficulty levels, that is, easy, moderate, and difficult. Overall, the test

has a mean difficulty value of 0.52, which indicates that the test is moderately difficult. Item discrimination values of the test items vary between 0.61 and 0.30. The total test has an item discrimination value of 0.42. Apparently, the test can distinguish students at a high level. Table 3 highlights the number of items, mean item difficulty, and item discrimination values for CT sub-tests.

Table 3. *Number of Items, Mean Item Difficulty and Item Discrimination Values for CT Test*

Sub-tests	Item Number	Mean Item Difficulty	Mean Item Discrimination
Inference	10	0.51	0.35
Evaluatingarguments	8	0.63	0.41
Deduction	11	0.49	0.49
Recognizingassumptions	12	0.45	0.41
Interpretation	10	0.55	0.43
Total	51	0.52	0.42

Table 3 shows the 51-item CT test with 10 items in the inference sub-test, 8 items in the sub-test of evaluating arguments, 11 items in the deduction sub-test, 12 items in the sub-test of recognizing assumptions, and 10 items in the interpretation sub-test. Arguably, the number of items is almost equally distributed across the sub-tests. Among all sub-tests, evaluating arguments has the highest mean difficulty level of 0.63, which means that this is the easiest sub-test in the test. This sub-test is followed by interpretation at a difficulty level of 0.55, inference at 0.51, deduction at 0.49, and recognizing assumptions at 0.45. With a mean item discrimination value of 0.49, deduction sub-test distinguishes the students the most in the test. This sub-test is followed by interpretation at 0.43, evaluating arguments at 0.41, recognizing assumptions at 0.41, and inference at 0.35.

For the criterion-related validity of the 51-item CT test of which content, face, and construct validities were achieved, the results of Adolescent Decision-Making Questionnaire and the CT test performed 2 weeks apart were compared with Kruskal–Wallis *H*-test due to non-normal distribution of data ($p< 0.05$), which are provided in Table 4.

Table 4.*Results of Kruskal-Wallis H Test Performed to See Whether Students' CT Scores Differed by Decision-Making Styles*

	Groups	*N*	*Mean Rank*	X^2	*sd*	*p*	*Difference*
Critical Thinking	Vigilance	100	78.60	23.288	3	.000	A-B A-C A-D
	Complacency	8	37.06				
	Panic	21	49.57				
	Cop out	8	32				

A: Vigilance B: Complacency C: Panic D: Cop out

According to Table 4, students' CT scores differed significantly by their decision-making styles ($X^2_{(sd=3,\ n=137)}$=23.288; p<0.05). To see whether this difference was between vigilance and other decision-making styles, Mann Whitney-U test was performed, and results are shown in Table 5.

Table 5.*Results of Mann-Whitney U Test Performed to See between Which Groups the Difference was by Decision-Making Styles*

	Groups	N	Mean Rank	Sum of Ranks	U	Z	p
Critical Thinking	Vigilance	100	56.88	5688.00	162.000	-2.816	0.005
	Complacency	8	24.75	198.00			
Critical Thinking	Vigilance	100	56.88	5688.00	590.000	-3.171	0.002
	Panic	21	39.10	821.00			
Critical Thinking	Vigilance	100	56.88	5688.00	138.500	-3.091	0.002
	Cop out	8	21.81	174.50			

As shown by the results of Mann–Whitney *U*-test performed between vigilance and other decision-making styles, there was a statistically significant difference between CT test scores of the students with vigilance decision-making style and the students with complacency (U = 162.000; $p< 0.05$), panic (U = 590.000; $p< 0.05$), and cop-out (U = 138.500; $p< 0.05$) decision-making styles. Thus, the students who were cautious and picky when making a decision had significantly higher CT test scores than the students who acted complacently,

panicked, and tended to avoid taking responsibility when making a decision. Hence, the CT test provided consistent results compared with decision-making styles. Consequently, criterion-related validity of the CT test was achieved.

Table 6 presents the results of Spearman–Brown correlation calculated for the sub-tests and total test scores of the draft CT test and WGCTAperformed 2 weeks apart to provide additional evidence both for content (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2014) and for construct (Terzi, 2019) validities.

Table 6. *Spearman-Brown Correlation Values for Two Critical Thinking Instruments*

Sub-tests	WG Inference	WG Evaluating arguments	WG Deduction	WG Recognizing assumptions	WG Interpretation	WG Total
Inference	0.338**					
Evaluating arguments		0.317**				
Deduction			0.323**			
Recognizing assumptions				0.476**		
Interpretation					0.412**	
Total						0.486**

WG= Watson-Glaser Critical Thinking Appraisal; **p<0.01

Table 6 shows a moderate significant correlation found between the sub-tests and total test scores of CT test and WGCTA. Thus, such correlation with WGCTAthat is commonly used in the literature can be offered as additional evidence for the content and construct validities of the CT test developed in the research.

3.2. Findings on the Reliability Study

Table 7 presents the results for KR20 reliability coefficient and test–retest correlation coefficient calculated for the CT test sub-tests and the total test.

Table 7.*Results for KR20 Reliability Coefficient and Test-Retest Correlation Coefficient Calculated for the CT Test Sub-Tests and Total Test*

Sub-tests	KR20	Test-Retest
Inference	0.62	r=0.57 (n=59, p<0.01)
Evaluating arguments	0.62	r=0.60 (n=59, p<0.01)
Deduction	0.76	r=0.58 (n=59, p<0.01)
Recognizing assumptions	0.64	r=0.70 (n=59, p<0.01)
Interpretation	0.75	r=0.70 (n=59, p<0.01)
Total	0.87	r=0.70 (n=59, p<0.01)

KR20 internal consistency coefficients were found to be 0.87 for the total test and varied between 0.62 and 0.76 for the sub-tests. Moreover, the test was split into two equal halves, and the correlation between the two halves was calculated to be 0.84. The correlation calculated between the results of the two applications, which were performed 3 months apart, ranged from 0.57 to 0.70. The correlation coefficient for the total test was found to be 0.70. Based on the results for KR20 internal consistency coefficient, split-half method, and test–retest method, the CT test provides reliable measurements. The correlation between the sub-tests and total test scores of CT test was checked, and the results are presented in Table 8.

Table 8. *Correlation Values between Sub-Tests and Total Test Scores of CT Test*

Sub-tests	Inference	Evaluating arguments	Deduction	Recognizing assumptions	Interpretation	Total
Inference	1	0.25**	0.45**	0.31**	0.36**	0.71**
Evaluating arguments			0.24**	0.23**	0.25**	0.45**
Deduction				0.45**	0.41**	0.72**
Recognizing assumptions					0.43**	0.67**
Interpretation						0.75**
Total						1

**p<0,01

Table 8 shows the significant correlations found among all sub-tests of the CT test ($p< 0.01$). While there was a low correlation between the sub-test of evaluating arguments and other sub-tests, moderate correlations were observed among other sub-tests. Also, a significant correlation was found between the total test score and the scores of each sub-test ($p< 0.01$). The correlation between the total test score and the scores of all sub-tests but the sub-test of evaluating arguments was observed to be high. A moderate correlation was found between the score of evaluating arguments and the total test score.

In summary, after the content and face validities had been achieved with expert opinions and the table of specifications and the criterion-related validity had been achieved with the Adolescent Decision-Making Questionnaire, the item and reliability analyses performed for the 51-item CT test show that it can measure CT skills of high school students in a valid and reliable manner.

4. Conclusion, Discussion and Recommendations

This study aimed to develop "Critical Thinking Skill Test for High School Students" to measure CT skills of high school students. Content, face, criterion-related, and construct validities of the "Critical Thinking Skill Test for High School Students" were examined. For the reliability studies, its KR20 coefficient was calculated, and test–retest and split-half methods were utilized. For the CT test prepared based on the sub-skills of inference, evaluating arguments, deduction, recognizing assumptions, and interpretation, which are deemed to represent the CT skill (Watson &Glaser, 1994), content and face validities were achieved with expert opinions and the table of specifications.

Following the item difficulty and item discrimination analyses performed to test the construct validity, 34 items were omitted from the test, which was finalized with 51 items. The 51-item CT test has 10 items in the inference sub-test, 8 items in the sub-test of evaluating arguments, 11 items in the deduction sub-test, 12 items in the sub-test of recognizing assumptions, and 10 items in the interpretation sub-test. While mean item difficulty values of the sub-tests vary between 0.51 and 0.63, mean item discrimination values range from 0.35 to 0.49. The total test has a mean item difficulty value of 0.52 and a mean item discrimination value of 0.42. Arguably, the test is moderately difficult and can highly distinguish students. To provide additional evidence for content and construct validities, the correlation between the CT test and WGCTA commonly used in the literature was checked, and a moderate significant correlation was found between the sub-tests and total test scores of the two tests.

Associated with CT skill in the literature (Halpern, 2003; Moore, 2010; Norris & Ennis, 1989; Bailin, 1998), decision-making skill was used to test the criterion-related validity of the "Critical Thinking Skill Test for High School Students". It was assumed that students with higher scores of vigilance as a decision-making style in the Adolescent Decision-Making Questionnaire, which was used for the criterion-related validity, would have higher CT skills and students with higher scores of complacency, panic, and cop-out as decision-making styles in the said questionnaire would have lower CT skills. These assumptions were confirmed in the relevant analyses. The students who were cautious and picky when making a decision had significantly higher CT test scores than the students who acted complacently, panicked, and tended to avoid taking responsibility when making a decision. Notably, the CT test provided consistent results compared with decision-making styles.

KR20 reliability coefficients calculated for the sub-tests ranged from 0.62 to 0.75. A KR20 reliability coefficient of 0.87 was calculated for the total test. Moreover, a correlation coefficient of 0.84 was calculated with the split-half method. To test time invariance of the test, the correlation values calculated between the results of the two applications, which were performed 3 months apart, ranged from 0.57 to 0.70. The correlation coefficient for the total test is 0.70. Based on the results for KR20 internal consistency coefficient, split-half method, and test–retest method, one can argue that the CT test will yield reliable results.

In light of the content and face validity achieved with expert opinions and the table of specifications and the criterion-related validity achieved with the Adolescent Decision-Making Questionnaire, the item analyses and reliability analyses performed for the "Critical Thinking Skill Test for High School Students" indicate that the 51-item test will provide valid and reliable results in measuring CT skills of high school students. In the test, the 10-item inference sub-test is composed of three-choice questions, the 8-item sub-test of evaluating arguments includes two-choice questions, the 11-item deduction sub-test is composed of four-

choice questions, the 12-item sub-test of recognizing assumptions consists of two-choice questions, and the 10-item interpretation includes three- and four-choice questions. The duration for applying the 51-question "Critical Thinking Skill Test for High School Students" is about 40 minutes. The highest possible score in the test is 51, and the lowest possible score is 0. A score between 0-17 refers to low CT skill, a score between 18-35 refers to moderate CT skill, and a score between 36-51 refers to high CT skill.

The validity and reliability studies of the "Critical Thinking Skill Test for High School Students" were carried out with students in different types of high schools located in a province in the West Black Sea Region. Applying the test to high school students in the provinces of other geographical regions to replicate the validity and reliability studies will assumably provide supportive evidence for test's validity and reliability. Validity and reliability studies can also be performed for using the "Critical Thinking Skill Test for High School Students" at different educational levels.

5. References

Ay, Ş. &Akgöl, H. (2008). Critical thinking, gender, age and grade level. *Journal of Theoretical Educational Science, 1*(2), 65-75.

Aybek, B. &Çelik, M. (2007). Watson-Glaser eleştirel akıl yürütme gücü ölçeğinin (W-GEAYGÖ) üniversite ikinci, üçüncü ve dördüncü sınıf İngilizce bölümü öğretmen adayları üzerindeki güvenirlik çalışması. *Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi, 16*(1), 101-112.

Aybek, B. (2006). *The Effect of content and skill based critical thinking teaching on prospective teachers'disposition and level in critical thinking*[Doctoral dissertation]. Çukurova University, Adana.

Bailin, S. (1998). Critical thinking and drama education. *Research in Drama Education: The Journal of Applied Theatre and Performance, 3*(2), 145-153.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş.& Demirel, F. (2014). *Bilimsel araştırma yöntemleri* (16. edition). Pegem.

Chance, P. (1986). *Thinking in the classroom: A survey of programs*. Teachers College, Columbia University.

Crocker, L. &Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cronbach, L. J. (1984). *Essentials for psychological testing*. Harper & Row.

Çıkrıkçı, N. (1993). Watson-Glaser eleştirel akıl yürütme gücü ölçeğinin (form Y M) lise öğrencileri üzerindeki ön deneme uygulaması. *Ankara University Journal of Educational Sciences, 25*(2), 559-569.

Çıkrıkçı, N. (1996, June). *Eleştirel düşünme: Bir ölçme aracı ve bir araştırma*. III. Ulusal Psikolojik Danışma ve Rehberlik Kongresi, Adana.

Çokluk, Ö., Şekercioğlu, G. &Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları* (2.edition). Pegem.

Çolakkadıoğlu, O. (2012). The reliability and validity study of adolescent decision making questionnaire for the high school students. *Mustafa Kemal University Journal of Social Sciences Institute, 9*(19), 387-403.

Demir, M. K. (2006). *The research of fourth and fifth grade primary school students' critical thinking levels in social studies lessons according to different variables. Journal of Gazi Educational Faculty, 26*(3), 155-169.

Eğmir, E. &Ocak, G. (2017). Eleştirel düşünme öğretim programının öğrencilerin eleştirel düşünme becerisi ve özdeğerlendirme düzeylerine etkisi. *Karaelmas Journal of Educational Sciences,* 5, 138-156.

Ennis, R. H. &Weir, E. (1985). *The Ennis-Weir critical thinking essay test.* Pacific Grove, CA: Critical Thinking Press and software.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*(2), 44-48.

Ennis, R. H., Millman, J. &Tomko, T. N. (2005). *Cornell critical thinking tests Level X and Level Z manual* (5th edition). The Critical Thinking Co.

Epstein, R. L. &Kernberger, C. (2012). *Critical thinking*. Advanced Reasoning Forum.

Evcen, D. (2002). *Adaptation of Watson-Glaser critical thinking apprasial test (form s) to* [Unpublished master thesis]. Ankara University, Ankara.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction - executive summary - the Delphi report.* The California Academic Press.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement and relationship to critical thinking skill. *Journal of Informal Logic, 20*(1), 61-84.

Fisher, A. (2001). *Critical thinking: An introduction*. Cambridge University Press.

Floyd, F. J. &Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Gorsuch, R. L. (2014). *Factor analysis*. Routledge.

Gülveren, H. (2007). *Critical thinking skills of education faculty students and factors influencing critical thinking skills* [Doctoral dissertation]. Dokuz Eylül University, İzmir.

Halpern, D. (2003). *Thought & knowledge: An introduction to critical thinking*. Lawrence.

Jonassen, D. H. (2000). *Computers as mindtools for schools: Engaging critical thinking.* Prentice Hall.

Jones, E., Hoffman, S., Moore, L., Ratcliff, G., Tibbetts, S. &Click, B. (1995). *National assessment of college student learning: Identifying the college graduates' essential skills in writing, speech and listening and critical thinking.* National Center for Educational Statistics.

Judge, B., Jones, P. &McCreery, E. (2009). *Critical thinking skills for education students*. Learning Matters.

Kan, A. (2011). Ölçme aracı geliştirme. S. Tekindal (Ed.). *Eğitimde ölçme ve değerlendirme* içinde (p. 239-276). Pegem.

Kennedy, M., Fisher, M. B. &Ennis, R. H. (1991). Critical thinking: Literature review and needed research. In L. Idol & B. Fly Jones (Eds.), *Educational values and cognitive instruction: Implications for reform* (pp. 11-40). Lawrence Erlbaum.

Kline, R. B. (2010). *Principles and practice of structural equation modeling*. Guilford Press.

Koç, C. (2007). *The effects of active learning on reading comprehension, critical thinking and classroom interaction* [Doctoral dissertation]. Dokuz Eylül University, İzmir.

Kurnaz, A. (2007). *Effects of skill and content-based critical thinking training on students' critical thinking skills, achievement and attitudes in the fifth grade course of social knowledge of primary school* [Doctoral dissertation]. Selçuk University, Konya.

Lewis, A. &Smith, D. (1993). Defining higher order thinking. *Theory into Practice, 32*(3), 131-137.

Mason, M. (2008). *Critical thinking and learning*. Blackwell Publishing.

Masterman, L. (1985). *Teaching the media*. Comedia Publishing Group.

Mecit, Ö. (2006). *The effect of 7E learning cycle model on the improvement of fifth grade students' critical thinking skills* [Doctoral dissertation]. Middle East Technical University, Ankara.

Norris, S. P. &Ennis, R. H. (1989). *Evaluating critical thinking*. Critical Thinking Press and Software.

Open Society Institute. (2021). *Media literacy index*. Access link: https://osis.bg/?p=3750&lang=en.

Özçelik, D. A. (2013). *Test hazırlama kılavuzu*. Pegem.

Pascarella, E. T. &Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. Jossey-Bass.

Paul, R., Binker, A. J. A., Jensen, K. &Kreklau, H. (1990). *Critical thinking handbook: 4th.-6th grades a guide for remodellingg lesson plans in langıage, arts, social studies & science*. Foundation for Critical Thinking, Sonoma State University.

Pérez Tornero, J. M. &Varis, T. (2010). *Media literacy and new humanism*. Moscow: UNESCO Institute for Information Technologies in Education.

Presseisen, B. Z. (1985). *Thinking skills throughout the K-12 curriculum: A conceptual design*. Philadelphia: Research for Better Schools.

Shipman, V. (1983). *New Jersey test of reasoning skills*. Upper Montclair NJ: IAPC, Test Division, Montclair State College.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences*. Routledge.

Swartz, R. J. &Parks, D. (1994). *Infusing the teaching of critical and creative thinking in elementary instruction.* Critical Thinking Books & Software.

Şenturan, L. (2006). *Critical thinking in nursing students* [Doctoral dissertation]. Marmara University, İstanbul.

Tabachnick, B. G. &L. S. Fidell. (2012). *Using multivariate statistics* (6. edition). Pearson.

Tekin, H. (2019). *Eğitimde ölçme ve değerlendirme* (27. edition). Yargı Yayınları.

Terzi, R. (2019). Nicel veri toplama teknikleri. S. Şen & İ. Yıldırım (Eds.), *Eğitimde araştırma yöntemleri* içinde (p.357-382). Nobel.

Tufan, D. (2008). *Critical thinking skills of prospective teachers: foreign language education case at the Middle East Technical University* [Unpublished master thesis]. Middle East Technical University, Ankara.

Turgut M. F. (1995). *Eğitimde ölçme ve değerlendirme metodları*. Yargıcı Matbaası.

Wagner, T. (2010). *The global achievement gap*. Basic Books.

Watson, G. & Glaser, M. E. (1964). *Watson-Glaser critical thinking appraisal manual*. Harcourt, Brace & World Inc.

Watson, G. &Glaser, M. E. (1994). *Watson-Glaser critical thinking appraisal form S manual*. San Antonio: The Psychological Corporation.

We Are Social. (2020). *Digital 2020*. Access link: https://wearesocial.com/digital-2020.

Wood, D. (1998). *Understanding children's worlds. How children think and learn: The social contexts of cognitive development* (2. edition). Blackwell Publishing.

Yıldırım, B. (2010). *The effect of skill based critical thinking education on the development of critical thinking in nurse students* [Doctoral dissertation]. Ege University, İzmir.