International Journal of Assessment Tools in Education
2022, Vol. 9, No. 3, 741–771
https://doi.org/10.21449/ijate.1169977
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of the article.

Pamukkale critical thinking skill scale: a validity and reliability study

Erdinc Duru[1], **Sevgi Ozgungor**[1], **Ozen Yildirim**[1,*], **Asuman Duatepe-Paksu**[2], **Sibel Duru**[1]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye
[2]Pamukkale University, Faculty of Education, Department of Mathematics and Science Education, Denizli, Türkiye

ARTICLE HISTORY

Received: Apr. 04, 2022
Accepted: Sep. 01, 2022

Keywords:
Critical thinking,
Test development,
University students,
Validity and reliability.

Abstract: The aim of this study is to develop a valid and reliable measurement tool that measures critical thinking skills of university students. Pamukkale Critical Thinking Skills Scale was developed as two separate forms; multiple choice and open-ended. The validity and reliability studies of the multiple-choice form were constructed on two different theoretical frameworks as classical test theory and item-response theory. According to classical test theory, exploratory and confirmatory factor analyses were performed, to item-response theory, the Generalized Partial Credit Model (GPCM) for one-dimensional and multi-category scales was tested for the construct validity of the multiple-choice form of the scale. Analysis results supported the unidimensional structure of the scale. The reliability analyzes showed that the internal consistency coefficient of the scale and the item-total correlation values were high enough. The test-retest analysis results supported that the scale shows stability over time regarding the field it measures. The results of the item-response theory-based analysis also showed that the scale met the item-model fit assumptions. In the evaluation of the open-ended form of the scale, a rubric was used. Several studies were conducted on the validity and reliability of the open-ended form, and the results of the analysis provided psychometric support for the validity and reliability. As a result, Pamukkale Critical Thinking Skills Scale, which was developed in two forms, is a valid and reliable measurement tool to measure critical thinking skills of university students. The findings were discussed in the light of the literature and some suggestions were given.

1. INTRODUCTION

All living things on earth have genus specific biological and cognitive resources that enable them to adapt to the world (Tolman, 1932, p. 374). The main skill that distinguishes humans in terms of these resources is thinking. Thinking as a core concept for all cognitive actions of human beings includes many important sub-processes. In this sense, the literature distinguishes among different high-level thinking processes such as reflective thinking, creative thinking, and critical thinking. Reflective thinking, like critical thinking conceptualized by Dewey (1933) and

*CONTACT: Ozen Yildirim ozen19@gmail.com Pamukkale University, Faculty of Education, Department of Measurement and Assessment, Denizli, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

often used interchangeably, can be defined as the process of creating a new understanding that makes the meaning and importance of the phenomenon apparent through processing the phenomenon and related intellectual experience analytically (Boyd & Fales, 1983). In other words, reflective thinking includes the process of making judgments for controlling and improving the learning process by actively thinking about what is known, what is lacking and how the discrepancy can be eliminated (Dewey, 1933). In the most general sense, creative thinking is defined as the ability to create new products (Parkhurst, 1999) by offering solutions or perspectives that have not been offered yet. On the other hand, critical thinking, referred by Paul (2005) as thinking about thinking, is defined as the evaluation and meaning making process (Mazer et al., 2007) of identifying the main themes and assumptions behind the claims presented in light of reasons and evidence independently of the effects of current prejudice and past experiences (Paul & Elder, 2001), discovering relationships, drawing conclusions based on existing evidence and considering whether these conclusions are valid based on the evidence (Pascarella & Terezini, 1991). Although each one of the aforementioned thinking processes is important in carrying out the daily life's actions and tasks which are getting more complex, diverse and requiring multi-dimensional perspective day by day, the critical thinking is no longer just an ability with extra advantage, rather, as the base of all other thinking processes (Paul & Nosich, 1991; Ruggiero, 1990), it has become an indispensable perquisite for the adaptation to today's world where information is temporary, intense and often misleading.As a matter of fact, in today's world, defined as the information age by many researchers in recent years, there has been numerous calls for critical thinking to be an indispensable part of the educational process (Facione, 2015; Lipman, 1988; Siegel, 1988; Uzuntiryaki & Capa-Aydin, 2013) as a necessary skill needed in every aspect of life including workplace performance and leadership skills (Flores et al., 2012), crisis prevention management, which has become more important under the threat of global warming (Comfort, 2007) and ensuring the continuation and preservation of democracy through knowledge management as a citizen under the information bombardment (Rezaee et al., 2012).

In spite of the existing need and it has a deep-rooted history dating back to Socrates (Bailey & Mentz, 2015), a common conceptualization that can provide scientific understanding and consistency in the literature has only emerged as late as 1990's as a result of the Delphi project (Mpofu & Maphalala, 2017). Awareness of the importance of critical thinking increased due to the regained popularity of the 1980's educational approach emphasizing the importance of inquiry-based high-level thinking skills where students are the main actors in the education process (Facione, 1990a). As a result of this awareness, The Delphi project was initiated to create a consensus-based conceptualization that could be used in critical thinking teaching by means of a holistic definition (Facione, 1990a). Within the scope of the project, 46 experts who are known for their contributions to the field from philosophy, education, social and natural sciences formed a committee and worked for two years to determine what critical thinking is and its most important components, skills, and related behaviors. After intense exchanges among prominent figures in the field such as Dave Ellis, Richard Paul, and Peter A. Facione, the committee determined that critical thinking has two inseparable components: critical thinking skills and critical thinking dispositions. Critical thinking skills further consist of interpretation, analysis, evaluation, inference, explanation, and self-regulation sub-skills. Interpretation skills, which are defined as understanding the content and importance of the text by critically considering the material at hand, independently of their own subjective thoughts, include cognitive skills such as recognizing the problem, defining it objectively, defining the content in one's own words, and defining the author's point of view. Analyzing is to identify the inferential relationships between elements by identifying the ideas and arguments presented in the content. In this framework, analysis includes actions such as finding the source of the claims in the content, identifying the similarities and differences between different options. Evaluation

includes skills such as deciding based on the credibility of a content to determine the reliability of the judgment, belief, decision, or ideas presented in the content and the reasons presented for these ideas, whether the evidence presented sufficiently supports the conclusion reached, or whether the reason, judgment or ideas put forward are within the framework of logic, existing situation, and evidence. Inference involves reasoning and drawing conclusions by questioning presented arguments and assumptions in the light of available evidence. In this framework, creating a consistent content-based synthesis, predicting the next step, and identifying possible outcomes are among the examples of making inferences. Explanation, which is another sub-dimension of critical thinking skills, is the ability to present the content as a coherent whole by synthesizing the ideas reached in the critical thinking process and is exemplified by actions such as presenting criteria that reflect the logic behind the decisions reached, turning the content into graphics. Self-regulation skill, which consists of self-testing and correction sub-skills, includes behaviors such as examining content objectively, reviewing previous decisions and ideas, referencing objective sources to be sure, and rearranging when erroneous inferences are noticed. In other words, self-regulation is the process of regulating one's own critical thinking process by critically addressing it. In addition to these sub-skills, Paul and Elder (2002) distinguish between weak critical thinking, which includes an objective analysis of the individual's content and is characterized as external to the individual, and strong critical thinking skills, which also includes the individual's monitoring of their own cognitive processes (p.38). Paul and Elder state that individuals with strong critical thinking listen to others even when they have completely different ideas from their own, try to understand by valuing their perspectives, and are able to change their own perspectives based on others' rationality. In other words, strong critical thinking also includes creating an objective reality by listening to others and evaluating events beyond their own personal needs within the framework of all other perspectives on the situation. This type of evaluation enables individuals to fully understand the others by putting themselves in the shoes of others and thus to develop a holistic understanding including the thinking of others and the underlying logic of this thought. For this reason, it can be argued that an important last sub-dimension of critical thinking is perspective taking, although it is not included in the Delphi project. In this context, perspective taking is the ability to approach the content from the perspective of others to express ideas based on the synthesis of different perspectives (Carpendale & Lewis, 2006) and to develop an original perspective.

One of the common deductions of the experts involved in the Delphi project is that these skills can be taught by training. As a matter of fact, the literature supports this inference. Bensley et al., (2010) reported that a significant increase was observed in the critical thinking scores of the psychology program students who took the research methods course when they received training on the critical thinking process in the first three weeks of the course, but the critical thinking scores of the students who did not receive this type of training did not change. Similarly, Cisneros (2009) stated that the critical thinking scores of the pharmacy students who did not receive any explicit critical thinking training in the study did not improve throughout the year, even though they had above the average scores compared to what is reported in the literature. On the other hand, there are many studies showing the positive effects of critical thinking activities when critical thinking is clearly instructed or when these activities are presented as a natural part of the course (see Abrami et al., 2008; Huber & Kuncel, 2016; Marin & Halpern, 2011; Mpofu & Maphalala, 2017; Msila, 2014; Sahool & Mohammed, 2018; Puig et al., 2019, for more detail).

The general conclusion drawn when the findings of these studies are taken together is that critical thinking can be supported by experience, strategic information and practice (Snyder & Snyder, 2008) gained through especially a teaching process that includes questions encouraging the analysis and evaluations of the claims behind the idea and arguments within the scope of

healthy skepticism (Browne & Freeman, 2000). At the same time, it was pointed out that the university life provides valuable experiences in the development of critical thinking (Huber & Kuncel, 2016; Pascarella & Terenzini, 2005), which naturally brings together many students with different cultural and life experiences together, and where discussion and analysis is supported more than previous educational experiences.

In many of the higher education institutions in Western societies, different methods are applied to support critical thinking skills as a result of this awareness (eg, Dumitru, et al., 2018). In Turkey, interest in critical thinking has increased recently, and although most of the studies have been carried out in the field of educational sciences (Batur & Özcan, 2020), the number of studies on critical thinking skills in different fields, especially health, business and economics, continues to increase. Most of these studies aim to determine the current situation (Batur & Özcan, 2020) and evaluate the competence levels of individuals' critical thinking skills. These studies indicate that university students' critical thinking skills are at medium or low level (e.g., Doğanay et al., 2007; Özmen, 2008). Another noteworthy point is that existing studies in the literature mostly use the concepts of skill and dispositions interchangeably and ignore the conceptual nuances between the two.

The critical thinking disposition, which constitutes the curiosity and motivation necessary for an individual to think critically, expresses the tendency or willingness of the individual to critical thinking skills such as questioning, thinking of alternatives and searching for evidence (Facione, 1990a). Facione identified seven critical thinking dispositions: analyticity, truth-seeking, self-confidence, maturity, open-mindedness, systematicity, and inquisitiveness. Although critical thinking dispositions are useful in predicting critical thinking skills as an integral part of critical thinking skills (Facione, 1997), unlike skills measured based on performance, it expresses a tendency to critical thinking and is measured through self-reports based on subjective evaluations. However, as mentioned above, critical thinking skills include performance based on deep processing of content through cognitive actions such as interpretation, analysis, evaluation, inference and explanation. Therefore, these skills could be measured only through testing the participant's ability to apply these skills instead of subjective evaluations of a person’s motivation to critically think.

Despite this distinction, the studies conducted in Turkey mostly use skills and tendencies synonymously, and dispositions are often tested in the evaluation of programs that claim to support critical thinking skills (e.g., Atay, Ekim, Gökkaya, & Sağım, 2009; Güçlü & Evcili, 2021; Nalçalı, et al., 2016; Özmen, 2008). In his comprehensive study that analyzed the historical development of critical thinking measurement tools in Turkey, Doğan (2013) stated that the psychometric properties of the scales for measuring skills have more psychometric issues compared to those measure dispositions. He also stressed the inadequacy of measurement tools based on adaptation studies as well as the need for the national psychometrically strong scale development studies.

Despite these apparent differences, an important reason why these two terms are used interchangeably is the limitations regarding the availability of a valid and reliable scale adapted to Turkish culture to measure critical thinking skills in the adult population. A series of tests have been developed in the literature to measure critical thinking. The most widely used of these tests in the literature are the Watson-Glaser Critical Reasoning Scale (WGCTA- Watson-Glaser Critical Thinking Apprasial, Watson & Glaser, 1980), The California Critical Thinking Skills Test (Facione &Facione, 1992), Cornel Critical Thinking Test Level X and Level Z (Cornell Critical Thinking Test Level X- Level Z) (Ennis & Millman, 1985) and New Jersey Test of Reasoning Skills (Shipman, 1983). Despite the intense work in the literature and the development of many measurement tools, the debate about the validity and reliability levels of these tests continues, and the findings are that the psychometric levels of these tests are not

ideal, or the findings are inconsistent (Abrami et al., 2008). In Turkey, validity and reliability studies were conducted on only a few of these scales -Watson-Glaser Critical Reasoning Scale and the California Critical Thinking Skills Test and the Cornell Critical Thinking test, which was developed to measure the critical thinking skills of preschool children- and many studies reported psychometric properties that were far from ideal.

Ayberk and Çelik (2007) collected data from pre-service teachers and reported reliability coefficients values ranging from .10 to .35 for the subscales of Watson-Glaser Critical Reasoning Scale, where the reliability coefficient for the whole scale was only .38. They pointed out that these numbers were similar to the values of .29 and .53 obtained by Evcen and Çıkrıkçı-Demirtaşlı (2002). On the other hand, the only subscale of The California Critical Thinking Skills Test commonly used in Turkey is the one measuring dispositions. The sub-scale of critical thinking skills has been shown to have reasonable psychometric values in studies conducted abroad (Facione, 1990b; Facione, 1990c), and although these findings were supported across different cultures, there are also call for caution regarding the use of the scale. For example, Jacobs (1995) reported that although the reliability coefficients of the A and B forms of the scale were .56 and .59, the reliability coefficients of the subscales were as low as .14. Moreover, although the scale has been translated into Turkish, it continues to be a measurement tool with very low accessibility since it is subject to a practically an unreachable fee in Turkey's conditions and is not equally open to all researchers.

Today, although critical thinking skills are needed in all areas of life and have become a prerequisite for the healthy functioning of society, there is currently no accessible scale to measure students' critical thinking skills in university that prepare individuals for working and living conditions and hence expected to support critical thinking skills. However, the lack of an accessible scale that can measure the critical thinking skills of students in university environments where critical thinking opportunities and development potential are abundant, makes it difficult to monitor whether the required improvements are achieved as a result of the current educational experiences offered in higher education institutions. At best, the lack of a valid scale limits the research scope to making predictions about critical thinking skills through dispositions.

In the light of the literature above, the need for an economical, accessible, valid and reliable measurement tool developed in Turkish culture is evident. In this context, the main purpose of this study is to develop a valid and reliable measurement tool for measuring critical thinking skills of university students in the context of Turkish culture.

2. METHOD

2.1. Participants

In the research, the aim was to develop both multiple choice and open-ended forms of the critical thinking skill scale, and for this purpose, data were collected from students studying in the field of teaching in different age groups and different departments. The data were collected according to convenient sampling method from prospective teachers studying in the 1st, 2nd, 3rd and 4th grades of Pamukkale University Faculty of Education between the 2019-2021 academic years.

During the construction of the open-ended form, data were collected from the participants in order to develop the rubric and to determine the response distributions, and then 15 participants were asked to answer the scale again for the reliability analysis of the test.

The data for the multiple-choice form were collected from two different groups. First, data were collected from 355 participants and analyzes based on Exploratory Factor Analysis (EFA) and Item Response Theory (IRT) were conducted. 29% (103 people) of the participants are male

and 70% (251) are female. The average age of the participants is 20.75. The Table 1 gives the distribution of participants by grade level.

Table 1. *The distribution of participants by grade level for EFA.*

Grade Level	Frequencies (*f*)	Percentage (%)
First	168	47.323
Second	80	22.535
Third	34	9.577
Fourth	73	20.281
total	355	100.00

The majority of the sample (47.32%) consists of 1st year prospective teachers. While the distributions of the 2nd (22.53%) and 4th grades (20.28%) are close to each other, it is seen that there are at least (9.57) 3rd year prospective teachers in the sample. The distribution of the participants according to the departments is given Table 2.

Table 2. *The distribution of the participants by the departments.*

	Frequencies (*f*)	Percentage (%)
Mathematics and Science	82	23.119
Turkish and Social Studies	47	13.260
Foreign languages	94	26.478
Special education	44	12.651
Guidance and Psychological Counseling	88	24.507
Total	355	100.000

The distribution of the participants participating in the research in the fields of mathematics and science (23.12%), foreign languages (26.48%) and guidance and psychological counseling (24.51%) is close to each other. In addition, the rate of these fields is higher than the fields of Turkish and social studies (13.26%) and special education (12.65%).

The scale was applied to 156 participants for Confirmatory Factor Analysis (CFA), which is used to determine the construct validity of the multiple-choice test. 26.00% of the participants are male (40 people), 74.00% are female (116). The predominance of female students in education faculties is a reflection of the sampling in the research. The average age of the participants was calculated as 21.91. Table 3 gives the distribution of the participants according to their grade levels.

Table 3. *Distribution of participants by grade level for the CFA.*

Grade level	Frequencies (*f*)	Percentage (%)
Second	56	35.897
Third	70	44.871
Fourth	30	19.232
Total	156	100.000

36% of the participants are second graders, 45% are third graders and 19% are fourth graders. The number of fourth graders in the sample is less than the second and third grades. The distribution of the participants according to their departments is given in Table 4.

Table 4. The distribution of the participants by the departments.

	Frequencies (*f*)	Percentage (%)
Mathematics and Science	49	31.410
Foreign languages	45	28.846
Guidance and Psychological Counseling	62	39.743
Total	156	100.000

Data was collected from three different departments. The number of participants participating in mathematics and science (31%) and foreign languages (29%) is close to each other, while the number of participants participating in Guidance and Psychological Counseling (40%) is higher.

2.2. Data Collection

In the research process, the theoretical framework was decided by analyzing the literature and existing scales to determine the type of measurement tool used to measure critical thinking skills (see Doğan, 2013, for more detail). The existing scales developed abroad (Watson-Glaser Critical Reasoning Scale, California Critical Thinking Skills Test, Cornel Critical Thinking Test Level X and Level Z, New jersey Thinking Skills Test), as well as the Critical Thinking Skills Test developed in Turkey (Eğmir & Ocak, 2016) were mostly observed to be in multiple-choice test format and in the form of independent questions. It was decided to form open-ended questions based on the text in order to evaluate the respondent's behavior at different cognitive levels, given an existing situation in the presentation of the relevant structure.

Selecting the text is a critical process in terms of guiding the further steps of the research. At the first step of the writing process of the essay, the topic was determined. The text was selected based on its relatedness to real life so that it could capture the respondents' attention, its depth and its suitability for preparing questions to tap different cognitive levels. In addition, the prior knowledge of the respondents and the difficulty of the text were taken into account, as it may affect the reader's understanding (Mullis et al., 2009). Among the different topics suggested by the researchers, *vaccines and today's reflections* were chosen as the subject. In the text, speculations based on the relationship between vaccine and autism and possible side effects of the vaccine are mentioned. It is an informative compilation text created by bringing together the information from different sources. Two Turkish language experts examined the text in terms of the criteria and grammar mentioned above, and the text was finalized by making the relevant corrections. An example of a short paragraph from the text is given below.

> *"While developing technology provides many conveniences in our lives, it has also brought some discussions. One of these debates is whether the vaccines made to protect our children from diseases by strengthening the immune system are associated with autism or not. In the last 20 years, cases of autism in developed countries have increased dramatically. While the probability of autism in a child born in the United States in 1992 was one in 150, this number increased to one in 68 in 2004."*

Upon construction of the text, open ended questions were written in light of the cognitive processes of critical thinking proposed in the literature (eg, Ennis, 1991; Facione, 1990a; Irani et al., 2007; Lippman, 1988; Norris & Ennis, 1990; Watson & Glaser, 1980) and therefore were decided to develop around seven cognitive processes The cognitive processes that are considered in the preparation of the questions and their definitions are as follows.

Interpretation: Understanding and expressing the meaning or significance of a wide variety of experiences, situations, data, events, judgments, conventions, beliefs, rules, steps or criteria. Sub-skills are classification, inferring and clarifying meanings.

Analyzing: Identifying inferential relationships between phrases, questions, concepts, explanations, or different forms of expression intended to express belief, judgment, experience, reason, knowledge, or opinions. Sub-skills are examining ideas, identifying arguments, and analyzing arguments.

Evaluation: Determining the reliability of explanations or definitions or statements made about perceptions, experiences, situations, decisions, beliefs or opinions, as well as; evaluating the logical strength of inferential relationships between statements, definitions, questions, or other representations. Sub-skills are evaluation of claims and evaluation of arguments.

Inference: Identifying the elements necessary to reach a logical conclusion, forming assumptions and hypotheses correctly, considering relevant information, and revealing results obtained from statements, principles, evidence, ideas, beliefs, opinions, concepts, questions and other forms of representation. Sub-skills are questioning evidence and reasoning and drawing conclusions about alternatives.

Explanation: Presenting one's reasoning results in a convincing and coherent way means being able to look at the big picture. Sub-skills are determining conclusions, justifying the steps, and presenting the arguments.

Self-regulation: Applying the cognitive activities, the elements used in these activities, and especially the skills of analysis and evaluation from the perspectives of questioning, validation, validation or correction to one's own inferential decisions. Sub-skills are self-testing and self-correction.

Perspective taking: Bringing different perspectives together and establishing cognitive empathy. In this sense, it can be said that perspective taking is a form of cognitive empathy.

Detailed information about the content of cognitive processes is included in the handbook of the scale. In this structure, the assumption that cognitive processes progress from an easy structure to a more complex structure has been accepted.

A total of 10 questions were composed/written based on two interpretations, two analyses, one evaluation, one inference, one explanation, two self-regulations and one perspective taking, based on the criteria specified above. In order to see the clarity of the questions, a pilot application was made with a sample of 10 participants, and the participants were asked questions that they did not understand or had difficulties. When the data were examined, it was observed that the desired answers could not be obtained, especially in the perspective taking question. Later, this question was reconsidered and revised by the researchers. The questions were rearranged by taking the opinions of a total of five experts in the field of critical thinking and measurement and evaluation before the pilot implementation.

The open-ended form consists of 10 items. The scale was applied to 136 participants within the scope of the research in order to develop the Rubric used in the evaluation of the scale. Then, the answers given to each question were brought together separately and analyzed and grouped from the most correct answer to the wrong answer. The answers given were grouped between one point and five points.

Based on the data received from experts and participants during the process, it was decided to develop a multiple-choice form in order to reduce the scoring bias of the scale, to facilitate the scoring and to enable it to be answered in a shorter time, in other words, to increase its usefulness (Cohen et al.., 1992; Ebel, 1972). As with open-ended questions, multiple-choice questions have options ranging from one to five points. The highest score that can be obtained from the test is 50 and the lowest score is 5. If the respondent receives zero (blank), one or two points from one of these two questions, he is deemed to have received zero points from the remaining items. Answers in the remainder of the test are not scored. In this case, the student gets zero points in total. Even if the student has not answered any question correctly, he/she can get zero points. While creating the options of the multiple-choice form, attention was paid to the followings:

- ✓ Harmony with the root in terms of grammar and meaning,
- ✓ Similar lengths of the options,
- ✓ Compatibility of the closeness of the distractors to the correct answer and the planned difficulty level of the items,
- ✓ The use of participants common mistakes in distractors (Bilican, 2021).

In addition, while preparing the test, the followings were considered in the test order;

- ✓ not to not placed the correct answers of the items in a certain pattern,
- ✓ to leave a certain gap between the items, the item root and the options,
- ✓ the suitability of the number of selected options to the level of the respondent,
- ✓ the first items are suitable for the lower level of the cognitive level and the last items are suitable for the last level of the cognitive level,
- ✓ to put a directive informing the students at the beginning of the test (Haladyna, 1997).

After the questions and options were written, five different experts working in the field of critical thinking (two critical thinking, one language, two assessment and evaluation) were asked to examine the multiple-choice test by considering the table of specification of the scale. The final version of the scale was determined according to the feedback received. In order to determine the psychometric properties of the scale, an application was made on two different study groups of 355 and 156 participants. One of the points to be considered in the scoring of the multiple-choice test and the open-ended test is that the first two interpretation skill questions in the test are criterion items. If the respondent gets zero (blank), one or two points from one of these two questions, she/he is deemed to have received zero points from the remaining items. In other words, in order to get points from the whole scale, the respondent must not score less than 2 in the first two questions. Under the leadership of Bloom, one of the most significant names in the education literature, many contemporary education researchers have conceptualized thinking skills in a spectrum ranging from basic processes such as knowledge, understanding and comprehension to thinking processes such as higher-level analysis and synthesis based on these basic processes (e.g. Anderson & Krathwohl, 2001; Crockett, 2019; Dwyer et al., 2014). The common argument of these conceptualizations is that the inferences reached by the reader who cannot grasp what the content means correctly will be wrong, and therefore, low-level comprehension skills form the basis of critical thinking skills (Dwyer et al., 2014). As a matter of fact, Williams et al. (2003) showed that a program for the development of critical thinking skills did not cause an increase in the critical thinking scores of students with low academic skills despite having the same feedback and practical experiences as other students, in other words, those who already have problems in understanding the text demonstrated the need for additional support to develop critical thinking. In this framework, the scores obtained from the other items measuring high-level skills such as analysis and evaluation, which should be formed within the scope of this basic understanding, were not calculated by the students who did not correctly answer the first two questions about the comprehension level of the text, and the scores of these students regarding their critical thinking skills were recorded as low. In such a case, it is assumed that the participants do not have the ability to answer other questions correctly and guess the answers by chance.

The rubric development process for the open-ended form of the scale was reconsidered after the development of the multiple-choice test. It was decided to give score points to the whole answer given by the student to each question and a holistic rubric was prepared. For this purpose, the answers received from 136 participants were examined and scores were graded for each level from the highest to the lowest. It was deemed appropriate to make the scoring between one point and five points. Participants who did not answer the question or answered meaninglessly were given zero points. In addition, what is expected from the respondent for each success level is written with clear descriptive statements. Using participants' responses based on these definitions, possible examples are given. Open-ended questions and the developed holistic rubric were finalized by taking the opinions of three assessment and evaluation experts. In order to determine the scoring reliability of the rubric, the open-ended test was applied to a similar sample of 15 participants. Responses were scored by five

researchers and three independent experts. In order to determine the consistency between the scores, the intraclass correlation coefficient between the five researchers and between a randomly selected expert from the research group and three independent experts was examined. A sample item and rubric are given below.

LEVEL: INTERPRETATION (Classifying, inferring and clarifying meanings)		
QUESTION 1. What do you think is the best title for this text?		
Score	**Evaluation Criteria**	**Sample answers**
5	Reflects the main theme (content/scope/focus of the text) of the text in the title, Explanation: The title fully reflects the relationship between vaccines and autism, which constitutes the content of the text.	• Relationship/link between vaccines and autism • Vaccination and autism • Discussions on the Relationship Between Vaccination and Autism
4	Although includes the main argument(s)/discussions of the text in the title, narrows the scope partially. Explanation: While examining the main focus of the text (the relationship between vaccines and autism), the title narrows it down to imply a causal relationship.	• Effects of vaccine on autism • Is Vaccine a Cause of Autism? • Do vaccines really cause autism?
3	Mentions only one of the main points that constitute the text content in the title (ya da mentions only one of texts content's main points in the title) Explanation: Although the main focus of the text is the relationship between shot and autism, limits the content by mentioning either only shot or autism in the title. Or even though mentions both, narrows the scope of at least one to the degree it does not reflect the text anymore.	• Is the vaccine our friend or not? • Autism and Its Causes • Vaccine and its importance· • The relationship between triple vaccine combination and autism
2	Although the title emphasizes the focus/most important elements/main elements of the text, it narrows the scope causing significant misunderstanding. Explanation: Although it mentions vaccine and/or autism in the title, it uses expressions that cause misunderstanding in a way that cannot be excluded from the scope of the text. Sets an irrelevant title that does not reflect the scope of the text.	• Does the developing technology trigger autism? • Vaccine-Autism Theory or Technology and Neuropsychiatry • Autism and infectious diseases • Increase in Diseases and Vaccination • Are vaccines killing our children?
1	Explanation: It does not include any statement that will reflect the relationship between vaccine and autism, which is the main element of the text.	• Incorrect treatment and possible consequences • Can technology make us worse while improving it? • Severe consequences of unfair ignorance • Science and diseases

The open-ended form and the multiple-choice form were applied separately to different groups in the classroom environment under the supervision of the researchers. While it took 20-30 minutes to answer the open-ended form, it took 10-15 minutes to answer the multiple-choice form.

2.3. Data Analysis

2.3.1. *Validity and reliability analysis of critical thinking multiple choice form according to Classical Test Theory*

Exploratory Factor Analysis (EFA) was performed to test the construct validity of the scale and to identify the items that best revealed the construct. Principal Axis Factoring Method, one of the factor determination methods, was preferred in EFA. Before the analyses, the assumptions of the factor analysis were tested. Univariate and multivariate outliers and missing values were examined in the data collected from a total of 355 participants, and finally, analyzes were carried out with a sample of 336 participants. Since the number of missing values was low (less than 5%) (Bennett, 2001; Shaffer, 1999) no data imputation method was used and they were excluded from the sample. The correlations between the ten items in the scale are given in Table 5.

Table 5. *Inter-item correlation coefficients for EFA.*

Item no	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
I1	1.000**									
I2	.304**	1.000**								
I3	.496**	.281**	1.000**							
I4	.504**	.211**	.464**	1.000**						
I5	.544**	.290**	.447**	.544**	1.000**					
I6	.569**	.281**	.519**	.591**	.592**	1.000**				
I7	.627**	.307**	.577**	.578**	.630**	.705**	1.000**			
I8	572**	.314**	.524**	.534**	.577**	.682**	.768**	1.000**		
I9	.559**	.298**	.472**	.553**	.542**	.583**	.702**	.739**	1.000**	
I10	.540**	.322**	.487**	.515**	.585**	.596**	.659**	.636**	.585**	1.000**

**$p<0.001$

The correlation coefficients between the items vary between 0.211 and 0.768. Although the correlation of I2 with other items is observed to be somewhat low (0.211 to 0.322), there are significant correlations between the variables according to the result of the Barlett test, which tests the significance of the correlation matrix and the suitability of the data for analysis, the data set is suitable for analysis ($p<0.01$). Finally, the KMO (Kaiser Mayer Olkin) value, which gives information about the suitability of the sample size for each variable and the whole model, was calculated as 0.947. It means the number of samples (336) used in the analysis is sufficient for the analysis.

While deciding the number of factors in EFA, it was tested with a parallel analysis in addition to the analysis results. The number of factors obtained from the factor analysis and the number of factors suggested by the additional analysis were compared in the scatter plot. The proof of reliability of the scale was calculated with the Cronbach's Alpha reliability coefficient, which gives information about internal consistency, and it was examined item discrimination, based on item-total test correlation, and the difference between the lower and upper 27% groups.

Test-retest reliability was also tested in order to obtain additional information about the reliability (in terms of stability) of the scale. For this purpose, the multiple-choice form of the critical thinking scale was applied twice to a similar sample group of 35 participants, one month apart, and the correlation between the first and second applications of the students was examined. In addition, the difference between the pretest-posttest scores of the critical thinking variable were examined using the paired sample t-test, and it was determined whether the variable changed over time. SPSS 26 was used for EFA and reliability analysis and Jamovi 2.3 program was used for parallel analysis.

To test the construct validity of the scale, confirmatory factor analysis (CFA) was performed with the data set collected from a different sample (156 participants) at the last stage. Before

the analysis, univariate and multivariate outliers were tested and two data were excluded from the analysis. In addition, the multivariate normality assumption was tested using Mardia's skewness and kurtosis coefficients. As a result of the analysis, it was seen that the data set did not meet the multivariate normality assumption (χ^2=509 p<0.001). For this reason, Robust Maximum Likehood (MLR) was used as the estimation method in CFA. Before the analysis, the adequacy of the correlation coefficients between the variables was examined. Table 6 shows the correlation coefficients between the items used for CFA.

Table 6. *Inter-item correlation coefficients for Table CFA.*

Item No	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
I1	1.000									
I2	.333**	1.000								
I3	.578**	.217**	1.000							
I4	.578**	.128	.430**	1.000						
I5	.597**	.166*	.392**	.570**	1.000					
I6	.632**	.197*	.558**	.615**	.628**	1.000				
I7	.671**	.187*	.581**	.591**	.574**	.745**	1.000			
I8	.645**	.184*	.492**	.530**	.546**	.699**	.776**	1.000		
I9	.626**	.172*	.444**	.576**	.488**	.599**	.728**	.782**	1.000	
I10	.588**	.203**	.461**	.546**	.614**	.679**	.675**	.621**	.618**	1.000

**p<0.001, *p<0.05

Except for I2 in the scale, correlations between items vary between 0.392 and 0.782. Although correlations between I2 and other items were significant at the 0.05 level, only one value (rm2-m4=0.128, p>0.01) was not significant. When the distribution of participants' answers to the I2 item was examined, it was determined that 73% (116) of 156 participants had the most correct answer, and the distribution was less than the other options. This may indicate a problem regarding the distinctiveness of the item. It was observed that the item measures the "interpretational behavior", which is an important step of critical thinking, and no problems were encountered in its writing or in the process of understanding the options. Due to the significant correlations between I2 and other variables, it was decided by the researchers that the item should remain on the scale. The decision of whether the item remains on the scale was decided according to the results of the CFA. Jamovi 2.3 program was used for CFA analysis.

2.3.2. *Validity and reliability of critical thinking multiple choice form according to Item Response Theory*

Measurement tools can be developed based on different theories, the validity and reliability proofs of the multiple-choice form of critical thinking based on Classical Test Theory (CTT) are given above. Traditionally, CTT is used in development tools. However, there are some limitations brought by the CTT, for example, the psychometric properties of a tool developed according to the CTT are affected by the characteristics of the individuals who answered the test. In another theory, Item Response Theory (IRT), item parameters can be evaluated independently of group characteristics and group characteristics can be evaluated independently of item sample (Hambleton & Swaminathan, 1985). For this reason, validity and reliability analyzes of the Critical Thinking Scale based on IRT were also tested. Due to the structure of the scale, parameter estimations were made using the Generalized Partial Credit Model (GPCM) for one-dimensional and multi-category scales. GPCM is a generalization of the 2-parameter logistic model (2PLM) used for items scored in two categories. For item discrimination *a* parameter and the difficulty *b* parameter is used which is one less than the number of categories. In addition, since GPCM is basically a logistic model, a value of 1.702 was used as the D scaling coefficient to approximate this model to the more mathematically complex *ogive* models. Analyzes were conducted on 336 participants. During the analysis,

catIRT tools (Aybek, 2021) and *mirt* (Chalmers, 2012) packages in R (R core team, 2022) were used in the creation of graphics. Before proceeding to the IRT analysis, the assumptions of unidimensionality, local independence and item model fit were tested. For the unidimensionality assumption, factor analytical methods were evaluated and the results of the EFA were examined, and for local independence, Yen's Q3 local independence statistic (Yen, 1993) was calculated. The critical cut-off point was accepted as 0.30 (Røe, Damsgård, Fors, & Anke, 2014). For item-model fit, RMSEA values were analyzed in the S_χ^2 statistic.

2.3.3. *Validity and reliability analysis of critical thinking open-ended form*

In order to ensure the reliability of the measurement tool, text and text-based questions were applied to 15 participants who were randomly selected and had sample characteristics, and then five experts who conducted the research scored the answers of the participants to each item based on rubric. Each item in the scale is scored multiple times. In determining the reliability of scores obtained from multiple-scored measurement tools, the inter-rater reliability coefficient can be determined by the intraclass correlation coefficient (ICC), which gives consistency between raters. As the evaluation of the ICC approaches 1.00, which can be interpreted as the evaluation of the correlation coefficient, the consistency between the raters increases, while the consistency decreases as it approaches 0.00. The suggestion of Portney and Watskins (2000) was taken into account in the evaluation of the coefficient obtained. Accordingly, when the sample size is less than 30 and the number of raters is less than 3, below 0.50 indicates weak reliability, 0.5-0.75 shows moderate reliability, 0.75-0.90 implies good reliability, and above 0.90 indicates excellent reliability.

Considering that the raters in the research group were together during both the development of the scale questions and the development of the rubrics, the consistency between the scores of three independent experts in the field of critical thinking and a randomly selected expert from the research group was evaluated by looking at the inte class correlation. The intraclass correlation coefficient was examined for both the item and the total scores obtained from the scale. 15 participants' responses were re-scored for one month by an expert selected from the research group in order to determine whether there was a difference between the scoring of the rater at two different times (intra-rater reliability). The correlations between the total scores given by the rater to each participant based on the first and second measurement results were examined. SPSS 26 program was used in the analysis.

In the research, it was tried to measure the same structure according to different measurement methods with the multiple choice and open-ended tests. The correlation between the scores obtained from these two scales in the study can also be considered as evidence for validity. Both tests were administered to 11 participants at different time intervals and the correlation between the scores was checked. Due to the small number of individuals, non-parametric Spearman Brown Rank Differences correlation analysis was performed.

3. FINDINGS

3.1. Validity and reliability findings of critical thinking multiple choice form according to Classical Test Theory

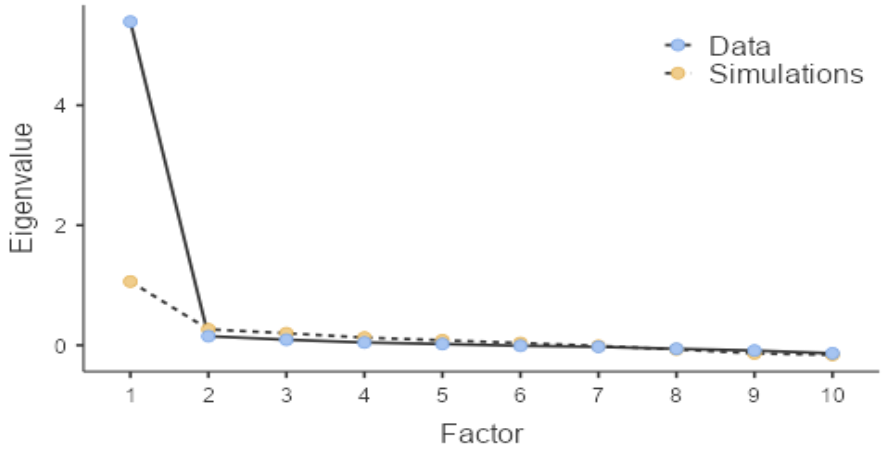
In the exploratory factor analysis, the contributions of ten items in the scale were examined and it was determined that except I2, they varied between 0.398 and 0.721. The contribution of I2 was calculated as 0.150, and the process was continued without removing the item from the analysis due to the reasons stated in the data analysis section. The explained total variance was examined to find the number of factors. Table 7 shows the explained total variance and eigenvalues.

Table 7. *Explained Eigenvalues and total variance.*

Factor	Total	Total %
I1	5.806	58.055
I2	.861	66.666
I3	.590	72.563
I4	.549	78.053
I5	.484	82.890
I6	.465	87.544
I7	.397	91.513
I8	.387	95.388
I9	.245	97.838
I10	.216	100.000

When Table 7 was examined, it was determined that there was only one factor (5.806) with an eigenvalue greater than 1 and this factor explained 58% of the total variance. It can be said that ten items were gathered under a single factor and explained more than half of the variance. In addition, when the parallel analysis scatter plot results based on the observed and expected values are examined, it is confirmed that ten items are grouped under a single factor.

Figure 1. *Parallel analysis scatter plot.*



Since the items were collected under a single factor, factor rotation was not performed. Finally, factor loading values were examined. Table 8 gives the factor loading values.

Table 8. *The factor loading values.*

Item No	Factor Loading
I1	.721
I2	.380
I3	.645
I4	.690
I5	.729
I6	.797
I7	.878
I8	.841
I9	.784
I10	.759

It is seen that the factor loads of the items are quite high (0.645-0.878). It was determined that only the factor load of I2 was 0.38, but this value was higher than the critical cut-off point of 0.30 (Tabachnick & Fidell, 2013). In addition, the RMSEA value of the model fit indices was

calculated as 0.047, which indicates a good fit (Browne & Cudek, 1993). Ten items show a single-factor structure.

After deciding on the number of factors, the reliability of the scale and the discrimination of the items were examined. The internal consistency coefficient of the scale was calculated with high reliability of 0.92. Although the number of items is small, it can be said that the scale is quite reliable in terms of internal consistency. The item-total test correlations are given in Table 9.

Table 9. *Item-total correlations and reliability values.*

Item No	r_{jx}	Cronbach's Alpha value when item is removed
I1	0.693**	0.910**
I2	0.367**	0.923**
I3	0.618**	0.913**
I4	0.664**	0.910**
I5	0.699**	0.908**
I6	0.762**	0.904**
I7	0.838**	0.900**
I8	0.804**	0.901**
I9	0.748**	0.905**
I10	0.728**	0.907**

**$p<0.001$

When the item-total test correlations were examined, the lowest 0.37 and the highest 0.838 correlation values were calculated. The item-test correlation value is above 0.30, it indicates that there is a sufficient relationship between the item and the construct to be measured (Tabachnick & Fidell, 2013). Although the item was removed from the model, it was observed that there was no significant change in the Cronbach Alpha value, and the Cronbach Alpha value calculated with ten items was quite high.

In order to support the validity and reliability of the scale, item discrimination was also calculated according to the lower and upper 27% groups. Table 10 gives item discrimination according to 27% lower-upper groups.

Table 10. *Independent samples t-test between lower-upper 27% groups.*

Item No	*t*-value
I1	11.858**
I2	8.822**
I3	11.859**
I4	10.871**
I5	17.189**
I6	13.098**
I7	14.779**
I8	16.589**
I9	17.127**
I10	17.008**

**$p<0.001$

The critical thinking scores of the participants in the lower and upper groups differ significantly for each item ($p<0.001$). The scores of participants with high critical thinking skills can be distinguished from the scores of participants with low critical thinking skills with the scale.

To calculate the reliability of the scale as stability, the test-retest reliability was checked and a correlation of 0.52 was calculated between the first and second application. There is a moderately significant positive correlation between the two measurements ($p<0.001$). The scores of the participants did not change between the first and second applications. In addition,

for the data obtained from these applications, the difference between the pretest-posttest scores of the critical thinking variable was not significant (p>.001). No change was observed in the participants' critical thinking skills during the process. Paired sample t-test results are given in Table 11.

Table 11. *Paired sample t-test result.*

Application	Mean	*N*	*SD*	Mean difference	*SD*	*t*	*df*	*p*
Pretest	38.085	35	5.537	0.771	5.041	0.905	34	0.372
Posttest	37.314	35	4.581					

CFA was performed to confirm the structure of the scale, which was found as a single factor. The overall goodness of fit values obtained when all items were added to the model and no modifications were made: χ^2/df=2.43 (Good), SRMR= 0.039 (Good), RMSEA= 0.095 (Poor), CFI=0.095 (Acceptable), TLI=0.94 (Low). According to Browne and Cudek (1993), a RMSEA value greater than 0.08 in the model indicates poor model-data fit. In addition, CFI and TLI values higher than 0.95 indicate good fit, while values between 0.90 and 0.95 indicate acceptable fit (Bentler, 1990). When the parameter estimations were examined, the standardized regression coefficients ranged between 0.628 and 0.884, while the standardized beta coefficient of I2 was significant at the 0.05 level (Beta= 0.248, p<0.05). While the variance rates explained by the items ranged between 0.39 and 0.78, I2 had the lowest explained variance (R^2= 0.06). The model should be revised according to the obtained values. According to these results, I2 was removed from the model and the analysis was repeated, the overall goodness of fit values obtained: χ^2/df= 2.72 (Good), SRMR= 0.037 (Good), RMSEA= 0.104 (Very Poor), CFI=0.095 (Acceptable), TLI Calculated as =0.94 (Low). It was observed that there was no change in the model fit values when I2 was added or removed from the model, and even when it was removed, the RMSEA value increased, and the model weakened.

Considering the cognitive process measured by I2, the researchers decided that I2 should remain in the scale, considering that I1 and I2 should be prerequisite items in scoring the scale, and that the prerequisite item should be measured with more than one item rather than a single item. In addition, instead of taking the average of all items in scale scoring, the validity of the presented answer was tested first. That is, although critical thinking is defined as a whole of multidimensional cognitive activities such as interpretation, understanding, and analysis (eg, Paul, 1990), the emergence of higher-level critical skills such as analysis and evaluation, and basic cognitive activities such as understanding and interpretation would not be possible without it. Therefore, in the present study, it is necessary to observe the cases where the questions I1 and I2, which measure the basic skills of understanding and interpretation, are answered incorrectly.

When the modifications are examined to increase the model fit, the error variances of I8 and I9, which measure self-regulation skills, are connected, the goodness of fit values increase (χ^2/df=62.8/34=1.85 (excellent), SRMR= 0.035 (Good), RMSEA= 0.073 (Acceptable) CFI=0.097 (Good), TLI=0.96 (Good)) and model data fit was observed. Since the distribution of the answers to I8 and I9 is similar and the items measure similar cognitive levels (self-regulation), this arrangement between errors was found appropriate by the researchers. The parameter values obtained from the model are given in Table 12.

When the standardized beta coefficients giving the relationships between the items and the factor were examined, it was observed that the lowest correlation was I2 (0.253) and the highest correlation was I7 (0.880). However, most of the items have a regression coefficient above 0.60.

When looking at the variance explaining the factor, while the contribution of I7 is the highest (0.774), the contribution of I2 is the least (0.064).

Table 12. *CFA model parameter estimation values.*

Item No	B	SH	β	z	R^2
I1	1.000	0.000	0.787		0.618
I2	0.224	0.094	0.253	2.37*	0.064
I3	1.184	0.099	0.638	11.89**	0.406
I4	1.090	0.088	0.704	12.30**	0.495
I5	1.386	0.112	0.707	12.30**	0.500
I6	1.540	0.114	0.846	13.45**	0.715
I7	1.585	0.115	0.880	13.70**	0.774
I8	1.603	0.117	0.823	13.68**	0.677
I9	1.570	0.118	0.773	13.23**	0.596
I10	1.598	0.122	0.778	13.04**	0.606

**$p<0.001$, *$p<0.05$

3.2. Validity and Reliability Analysis Findings of Critical Thinking Multiple Choice Test Based on Item Response Theory

When the EFA results, which were conducted to determine the unidimensionality of the critical thinking scale, it was determined that there was only one factor with an eigenvalue greater than 1 (5,806) and this factor explained 58% of the total variance. Accordingly, it was found that ten items were gathered under a single factor and explained more than half of the variance.

Violation of local independence may affect individual parameter estimates, reliability and validity estimates of the scale (Marais, 2009; Yen 1993). For this reason, the second assumption of the IRT, local independence, was tested and it was seen that all items were below the critical cut-off point (0.30) according to the Yen's Q3 local independence test and did not violate local independence. As the last assumption, item model fit was examined, and item calibrations were made according to GPCM. The S_χ^2 statistic for item concordance is given in Table 13.

Table 13. *Item fit indices.*

	S_χ^2	*df*	RMSEA
I1	28.620	27	0.013
I2	97.837	28	0.086
I3	78.936	55	0.036
I4	54.391	45	0.025
I5	75.403	47	0.042
I6	56.867	43	0.031
I7	45.416	35	0.030
I8	37.517	34	0.018
I9	62.882	41	0.040
I10	83.608	47	0.048

It was determined that the RMSEA values of nine items in the scale ranged between 0.013 and 0.048, and these items fit well with the model. The RMSEA value of I2 was calculated as 0.086. This item has low agreement with the model. A similar situation was observed in both the item discrimination and the contribution of the item to the model in the EFA and CFA analyzes, but it was decided to keep the item based on expert opinion.

After deciding on the model item fit, item parameters and standard errors of these parameters were calculated. The values of the parameters are given in Table 14.

Table 14. *Parameter values and standard errors of items according to GKPM.*

Item no	*a*	*b1 (0-1)*	*b2 (1-2)*	*b3 (2-3)*	*b4 (3-4)*	*b5 (4-5)*
I1	0.827 (0.114)	NA	-3.570 (0.505)	-1.235 (0.317)	-0.259 (0.279)	-1.549 (0.339)
I2	0.347 (0.064)	NA	NA	-5.473 (1.165)	2.457 (0.897)	-6.702 (1.406)
I3	0.848 (0.124)	-0.652 (0.267)	-0.064 (0.254)	0.013 (0.240)	0.310 (0.219)	0.809 (0.228)
I4	1.010 (0.131)	0.067 (0.343)	-1.815 (0.367)	0.588 (0.163)	0.960 (0.202)	1.096 (0.248)
I5	1.153 (0.177)	-0.612 (0.247)	-0.260 (0.231)	-0.315(0.202)	0.818 (0.190)	-0.129 (0.200)
I6	0.996 (0.147)	0.080 (0.392)	-0.490 (0.385)	-1.298 (0.360)	0.200 (0.174)	0.309 (0.166)
I7	1.399 (0.221)	0.380 (0.492)	-1.365 (0.470)	-0.856 (0.263)	-0.121 (0.141)	0.657 (0.123)
I8	1.603 (0.276)	-0.439 (0.268)	-0.530 (0.251)	0.053 (0.186)	-0.509 (0.194)	0.318 (0.108)
I9	1.868 (0.374)	-0.865 (0.178)	0.420 (0.194)	0.019 (0.195)	-0.427 (0.198)	0.112 (0.120)
I10	1.002 (0.169)	-0.531 (0.284)	0.261 (0.307)	-0.325 (0.312)	-0.239 (0.249)	-0.558 (0.240)

According to Table 14, it is observed that the discrimination parameters of the items (a) are close to 1.00. According to Baker (2001), 0.01-0.34 is considered very low, 0.35-0.64 low, 0.65-1.34 moderate, 1.35-1.69 high, and 1.70 and above very high. Item discrimination gives information about the power of the item to distinguish students according to their abilities. The higher the discrimination, the better the item can distinguish individuals according to the relevant structure. Accordingly, six items (I1, I3, I4, I5, I6, I10) have medium discrimination, 2 items (I7 and I8) have high discrimination, and one item (I9) has very high discrimination. The discrimination of I2 is low. The other predicted parameter is the "b_i" (option response function) parameter, which gives information about the item difficulty or the item response frequency. In GPCM, the number of option response functions is one less than the number of possible options. Since the scale was scored between 0 and 5, five alternative response functions were calculated. However, when the response pattern of I1 and I2 was examined, b1 and b2 of these items could not be calculated since there were no students who got zero points in M1 and no students who got zero and one points in I2. Option response parameters ranged from -6,720 to 2,457. When the b parameters are examined, it is seen that the b parameters of the items other than the 1st and 2nd items used as prerequisites include individuals with both low and high critical thinking levels. Item characteristic curves and item test information functions of ten items in the scale are given in Figure 2 and Figure 3.

When the item probability functions and item information functions are examined together, it is seen that I2 does not provide information for all ability levels. The item probability function of this item focused especially on two score categories. These categories are 2 (P2) and 5 (P5). Therefore, individuals below -2 ability level are more likely to get 3 points from this item, while individuals above -2 ability level are more likely to get 5 points from this item. Other score categories for this item could not be differentiated for different ability levels. On the other hand, I9 provides very high information especially for individuals between -2 and +2 skill levels.

Figure 2. *Item probability functions.*

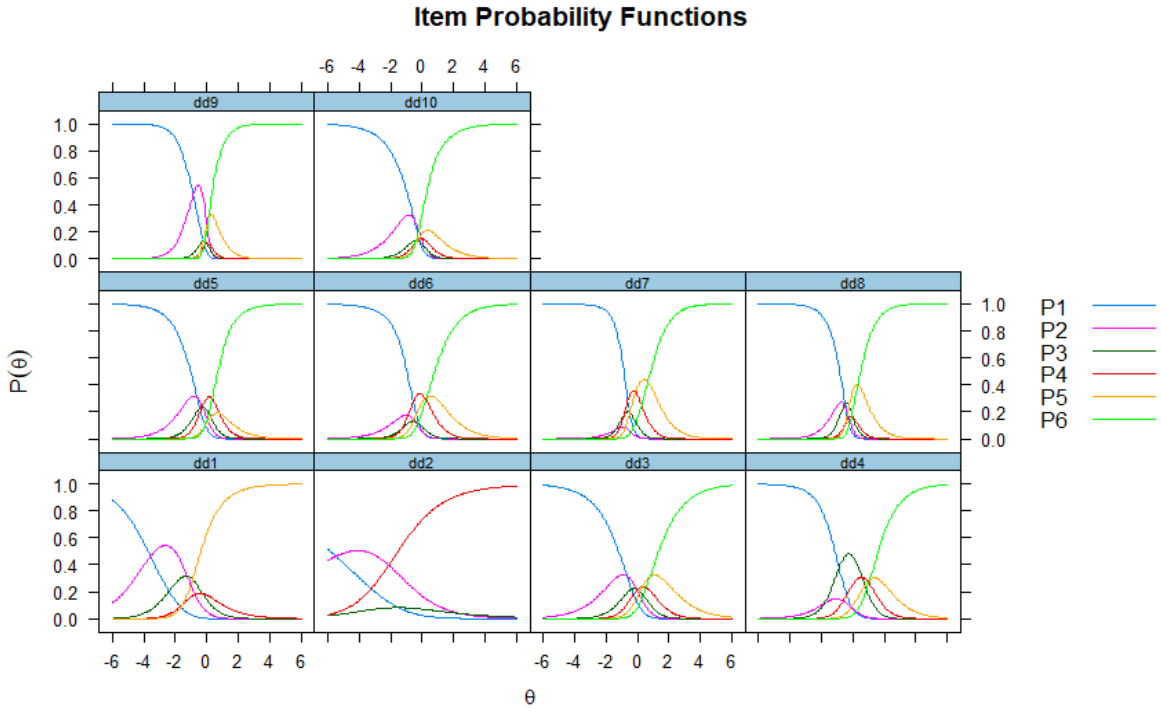
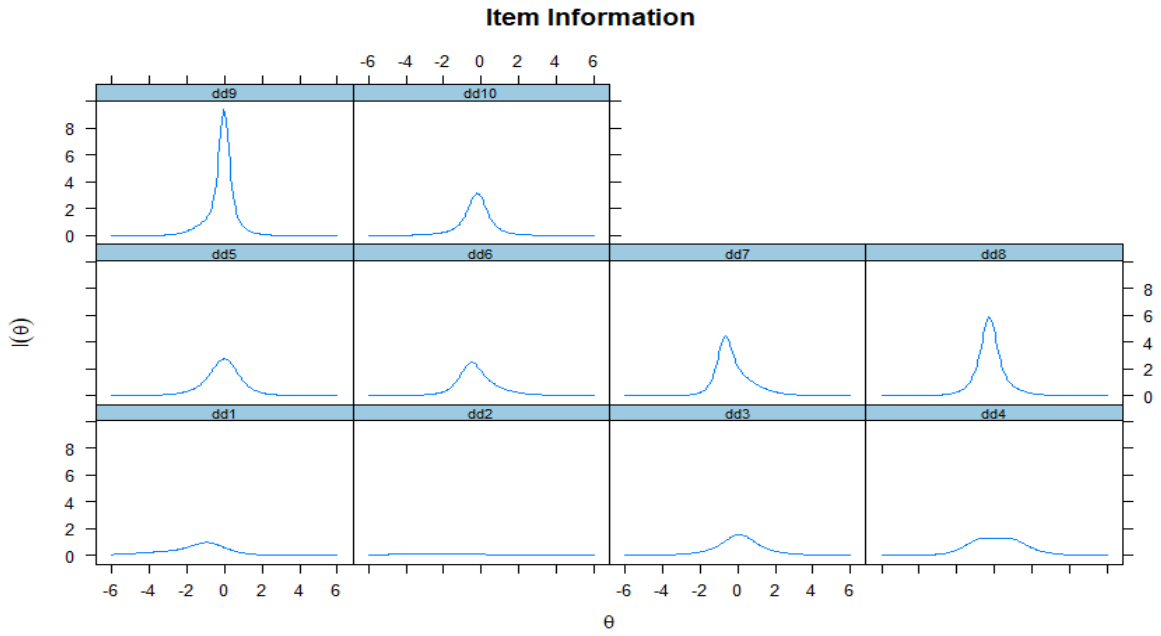
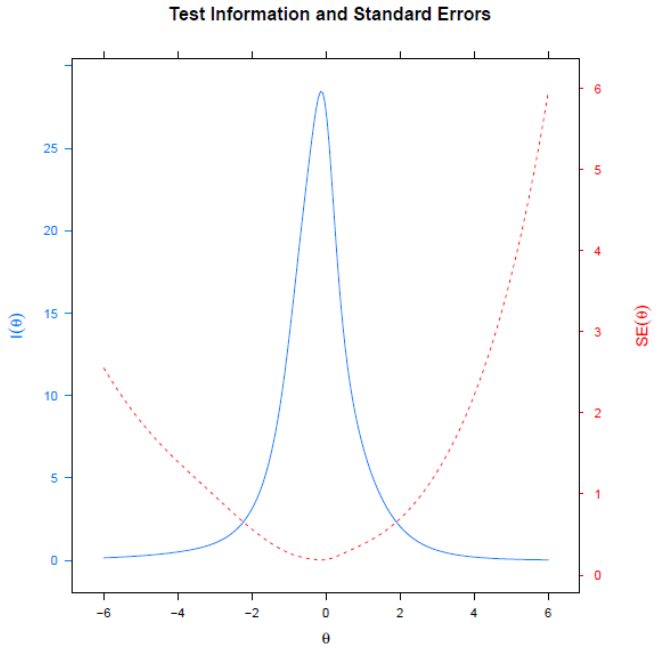


Figure3. *Item information.*



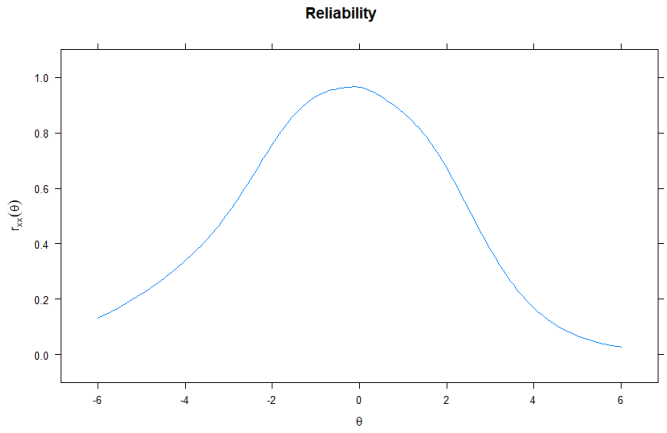
The test information function was evaluated (Figure 4), it was seen that the scale provided more information for individuals whose critical thinking levels were between -2 and +2, in other words, it distinguished these individuals better.

Figure 4. *Test information functions.*



The empirical reliability coefficient of the scale based on IRT was calculated as .91. In addition, when the reliability function obtained for all ability levels is examined, the scale measures with a reliability above .80 especially for individuals between -2 and +2 ability levels. Figure 5 shows the reliability of the test.

Figure 5. *Reliability of the test.*



3.3. Critical Thinking Open-Ended Form Holistic Rubric Validity and Reliability Findings

To ensure the validity steps of the holistic rubric, the steps described in detail in the data collection section were followed. Extra care has been taken to ensure inter-research coherence in arranging the definitions based on each question and possible participant responses to these definitions. In addition, the developed scoring tool was finalized by taking the opinions of experts in two critical thinking, one Turkish language and three measurement and evaluation fields and making necessary adjustments.

To determine the reliability of the scoring key, the consistency between the raters was examined. The correlation coefficients between the scores given by the five experts in the study to the answers of 15 participants are given in Table 15.

Table 15. *Intraclass correlation coefficient (five experts).*

Item no	Intraclass correlation coefficient
I1	0.992**
I2	0.956**
I3	0.926**
I4	0.993**
I5	0.994**
I6	0.974**
I7	0.993**
I8	0.968**
I9	0.991**
I10	0.985**
Total score	0.966**

**$p<0.001$

Table 15 displays that the intra-class correlation coefficients range between the raters according to the items vary between 0.956 and 0.992. It can be said that the consistency between raters is quite high. When the total scores given by the raters to 10 items were compared, it was observed that the correlation coefficient between the raters was again very high (0.966).

When the correlations between independent raters were examined in order to support the consistency between raters and to avoid bias, Table 16 correlation coefficients were obtained.

Table 16. *Intraclass correlation coefficient (four independent experts).*

Item no	Intraclass correlation coefficient
I1	0.983**
I2	0.841**
I3	0.628**
I4	0.962**
I5	0.746**
I6	0.956**
I7	0.832**
I8	0.956**
I9	0.876**
I10	0.877**
Total score	0.925**

**$p<0.001$

The consistency between the scores given by the four experts to each item varies between 0.628 and 0.962. While the consistency between the raters was quite high for nine items, it was observed that the consistency between raters was moderate (0.628) in I3. In addition, there is a high (0.925) consistency among the raters according to the total scores. The evidence obtained reveals that reliable scoring can be done with the holistic rubric developed in scoring responses.

The correlation coefficient obtained as a result of the same rater scoring 15 participants at different times was calculated as 0.765 ($p<0.001$). There is a positive high-significant relationship between the rater's first and second evaluations made with a one-month interval. This situation reveals that the scale also provides reliability against time.

In the evaluation of the same structure according to different measurement types, the correlation between the scores of 11 participants from multiple choice and open-ended tests varied between 0.461 and 0.658 for 5 raters. Accordingly, there was a positive moderate significant correlation

between the scores obtained from the multiple-choice test and the open-ended test (p<0.001). It is expected that the correlation coefficient between the same constructs will be high, but it should be considered that the answers to the multiple-choice test are more structured, and the objectivity is strong in the evaluation, whereas the bias stemming from the expert opinion in the evaluation of the answers to the open-ended test should be considered. Therefore, this may be the reason for the possible decrease in the correlation coefficient.

4. DISCUSSION, CONCLUSION and SUGGESTIONS

The main purpose of this study is to develop a measurement tool with high validity and reliability that measures critical thinking skills of university students. In this context, a series of studies were conducted on different participant groups. The results of the analysis support that the psychometric properties of the developed scale are acceptable and can be used to evaluate critical thinking skills of university students.

In accordance with the scale development procedures, first of all, the literature was searched and the existing scales in the literature were examined (Ennis & Millman, 1985; Facione & Facione, 1992; Shipman, 1983; Watson & Glaser, 1980). The existing scales are mostly in multiple choice test format and in the form of independent questions/ the literature mainly includes scales in multiple choice test format and in the form of independent questions. In addition, these scales do not have strong psychometric properties or that the findings are not supported by different research results (Abrami et al., 2008). A similar situation seems to be valid for a few scales adapted for use in our country. In these studies, psychometric properties that are far from expected regarding the adapted scales were reported (Ayberk & Çelik, 2007). Therefore, in order to capture the critical thinking potential; It is thought that it is necessary to use performance-based evaluation rather than self-reports, critical thinking consists of related abstract cognitive structures, and it is more appropriate to conduct a holistic evaluation by evaluating the cognitive level reactions of the participants while presenting a case to identify these structures. For this purpose, it was decided to develop two separate forms of the Pamukkale Critical Thinking Skills Scale, which is structured based on the selected text; multiple choice format and open-ended format. The validity and reliability studies of the multiple-choice critical thinking scale are based on two different theoretical frameworks: classical test theory and item-response theory. In the evaluation of the open-ended form of the scale, the developed "rubric" was used.

According to classical test theory, a series of analyzes were conducted to test the construct validity of the multiple-choice form of the critical thinking scale. Whether the partial correlations between the items and the correlation matrix were suitable for factor analysis were examined with the Kaiser-Meyer-Olkin (KMO) coefficient and the Barlett test (Fayers & Machin, 1995). The analyzes showed that the KMO value was high and the Barlett test result was significant. In order to determine the construct validity of the scale, factors with an eigenvalue above 1.00 according to Kaiser normalization were taken as the criteria. The findings showed that the items were collected on a single factor of 5,806 eigenvalues, which constituted 58% of the total variance. Considering that the variance explained in social sciences should be at least 40% and above (Stevens, 1992), the results of the analysis seem significant. When the items that make up the scale were analyzed in terms of factor loads, it was observed that the factor loads of the items ranged from .38 to .84. The fact that the factor loads are above .30 is considered important in terms of showing the high representativeness power of the items in the scale. Similarly, the break point of the graph supports that the breakout occurred after the first factor.

After the exploratory factor analysis, Confirmatory Factor Analysis (CFA) was performed on a different study group to confirm the single-factor structure of the scale. When all items were added to the model and no modifications were made, some of the general goodness of fit values

were higher than expected (χ^2/df=2.43, SRMR= .039, RMSEA= .95, CFI=.95, TLI=.94). For example, according to Browne and Cudek (1993), a RMSEA value greater than .08 in the model indicates poor model-data fit. When the parameter estimations were examined, the standardized regression coefficients ranged from .62 to .88, while the beta coefficient of I2 was low (.25) but significant at the .05 level. All items except for I2 have a beta coefficient over .60. According to the obtained values, the model should be revised, and some modifications should be made. According to these results, I2 was removed from the model and the analysis was repeated, but it was observed that there was no change in the model fit values when I2 was added or removed from the model and even RMSEA value increased, and the model weakened when I2 was removed. Considering the cognitive feature measured by I2, it was decided to keep I2 in the scale. In addition, some modifications were made to increase model compatibility. It was observed that when the error variances of I8 and I9 were connected, the goodness of fit values increased (χ^2/df=62.8/34=1.85, SRMR= 0.035, RMSEA= 0.073, CFI=0.097, TLI=0.96) and the model data fit increased. It can be said that this arrangement between error variances is appropriate since the distribution of the answers given by the students to I8 and I9 is similar, and the items measure similar cognitive characteristics (self-regulation). In summary, the results of EFA and CFA analysis support the one-dimensional structure of the scale.

Related to the reliability studies of the scale, the internal consistency was calculated with the Cronbach's Alpha reliability coefficient, the item-total test correlations were examined and the level of discrimination between the upper and lower groups of the items was examined, and the test-retest method was used to test the measurement stability. In the data analyzes, the internal consistency coefficient of the scale was calculated as .92. This result shows that the similarity of the items and the consistency of the responses to the items are high.

When the item-total correlations and correlation matrix of the scale were examined, it was observed that the correlation values ranged between .37 and .84. If the item-test correlation value is above .30, it indicates that there is a sufficient relationship between the item and the construct to be measured (Tabachnick & Fidell, 2013). According to these results, it can be said that the items of the scale are positively and significantly related to each other and the whole scale.

In order to support the validity and reliability of the scale, item discriminations were also calculated according to the lower and upper 27 % groups, and it was observed that the critical thinking scores of the participants in the lower and upper groups differed significantly for each item. According to these results, it can be said that the scale can significantly distinguish the scores of participants with high critical thinking skills from the scores of participants with low critical thinking skills.

The test-retest method was used to test the measurement stability. For this purpose, the scale was administered to the participants with an interval of three weeks, and the Pearson Correlation Coefficient between the two applications was found to be significant at the level of .52. The results of the analysis also showed that there was no significant change in the scores of the participants between the first and second applications. These results indicate that the scale shows stability over time regarding the behavioral domain it measures. In other words, no significant change was observed in participants' critical thinking skills over time.

Measurement tools can be developed based on different theories, the validity and reliability evidence of the multiple-choice form of critical thinking based on Classical Test Theory (CTT) are given above. In addition, validity and reliability analyzes of the Critical Thinking Scale based on IRT were also conducted. Before proceeding to the IRT analysis, the assumptions of unidimensionality, local independence and item model fit were tested. Considering the one-dimensional assumption of the theory, EFA and CFA results support that Pamukkale Critical Thinking Skills Scale is one-dimensional. In addition, local independence, the second

assumption of the ITC, was tested and it was seen that according to the Yen's Q3 local independence test, the critical cut-off point for all items was below .30 and did not violate local independence. As the final assumption, item model fit was examined, and item calibrations were examined. According to the results of the analysis, it was observed that the RMSEA values of most of the items in the scale ranged between .013 and .048. Only the RMSEA of I2 was slightly higher than expected (.086).

After deciding on the model item fit, the item parameters and the standard errors of these parameters were calculated, and it was observed that the discrimination parameters (a) of the items were close to 1.00. Accordingly, six items (I1, I3, I4, I5, I6, I10) have medium discrimination, 2 items (I7 and I8) have high discrimination, and one item (I9) has very high discrimination. The discrimination of I2 is low. Since the scale was scored between 0 and 5, five alternative response functions were calculated. When the b parameters were examined, it was seen that the b parameters of the items other than the 1st and 2nd items used as prerequisites in scoring included both individuals with low and high critical thinking levels.

When the alternative response functions and item information functions are examined together, it can be said that the scale provides more information for individuals whose critical thinking levels are between -2 and +2, in other words, it distinguishes these individuals. The reliability coefficient of the scale based on IRT was calculated as .91. In addition, when the reliability function obtained for all skill levels is examined, the scale measures with a reliability above .80 especially for individuals between -2 and +2 skill levels. As a result, IRT-based analysis results of the scale; unidimensionality, local independence, and item-model fit assumptions.

In addition, a number of studies were conducted on the validity and reliability of the open-ended form of the Pamukkale Critical Thinking Skills Scale, and the detailed steps given in the data collection section were followed in order to develop the rubric. In order to determine the reliability of the scoring key, the consistency between the raters was examined. According to the analysis results, the inter-class correlation coefficients between raters ranged from .95 to .99. When the total scores given by the raters to the 10 items were compared, it was observed that the correlation coefficient between the raters was again quite high (.97). These results can be considered as an indication that the rubric is well structured and therefore the consistency between raters is high. Correlations between independent raters were also examined to control for the possibility of inter-rater bias. For this, the evaluations of four experts were used. The consistency between the scores given by the four experts to each item varies between .62 and .96. In addition, it was observed that there was a very high (.92) consistency between the raters according to the total scores. The evidence obtained reveals that reliable scoring can be done with the holistic rubric developed in scoring participant responses. In addition, the correlation coefficient obtained as a result of the same rater scoring 15 participants at different times was calculated as .76. There is a highly significant positive correlation between the rater's first and second evaluations made one-month apart. This situation can be evaluated as an indication that the scale has reliability over time.

Finally, it was examined whether two separate scale forms developed to measure critical thinking skills could make similar evaluations. The results of the analysis showed that the correlation between multiple choice and open-ended tests scores of 11 participants was .46 and .65. According to these results, it can be said that there is a moderate positive correlation between the scores obtained from the multiple-choice test and the open-ended test. On the other hand, considering that the answers to the multiple-choice test are structured, the objectivity is strong in the evaluation, and that there may be some limitations in the evaluation of the answers to the open-ended test due to expert opinion, this result seems significant.

In summary, the main purpose of this study was to develop a valid and reliable measurement tool that measures critical thinking skills in university students. The results of the analysis

provided psychometric support that the measurement tool developed in two forms is valid and reliable and can be used to measure critical thinking skills of university students. Considering the limited number of measurement tools that measure critical thinking skills based on performance, it can be said that the study contributes to the literature (Abrami et al., 2008; Facione & Facione, 1992; Ennis & Millman, 1985; Shipman. 1983; Watson & Glaser, 1980). In addition, the study contributes to the literature in terms of conceptual perspective as well as scale forms. A new conceptual dimension called "Taking Perspective" was added to the existing critical thinking dimensions and this was supported by the findings of the research. As a meaningful component of critical thinking, perspective taking requires the individual to be able to both connect with the person, text, situation, or theme and stay objective by keeping a distance from them. In addition, different from the Delphi project, the operational measurement of the "Self-Regulation" skill and its inclusion as a basic component in the content of the developed scale can be considered as another contribution to the literature. Therefore, the conceptual framework of the study can form the basis for the structuring of educational programs in the processes of developing and teaching critical thinking, which is conceptualized as an important 21st century skill (Duru et al., 2020; Trilling & Fadel, 2009; Van Laar et al., 2019; Voogt & Roblin, 2012).

Similarly, the development of the Pamukkale Critical Thinking Skills Scale in two separate forms, open-ended and multiple-choice, is another important contribution to the literature. While the open-ended form allows to evaluate critical thinking skill in a holistic and performance-based manner, the use of multiple-choice form, free from chance factor, seems to be advantageous in terms of practical, economic, accessible and time, besides holistic evaluation. Therefore, it can be expected that the scale will help field experts and educators both in understanding the level of critical thinking skills of university students and in evaluating the contribution of curriculum and practices to the development of critical thinking skills. In addition, critical thinking is one of the higher-order thinking skills, and individuals with this potential can be considered qualified human resources. Therefore, the conceptual framework related to scale can contribute to policymakers in determining qualified human resources and creating, developing, and planning education policies for this resource. Finally, it can be said that the two most important features that distinguish the Pamukkale Critical Thinking Scale (PCTS) from similar scales in the literature are that it measures critical thinking skills on a performance-based way with a text and can evaluate the individual as a whole in terms of critical thinking skills.

In the light of the above explanations, the findings of this study should be evaluated within the framework of some limitations. First, in this study, the psychometric properties of the Pamukkale Critical Thinking Skills Scale were tested on the students of the Faculty of Education. Therefore, examining the psychometric properties of the scales on different study groups and in different universities may contribute to the validity and reliability of the scale and the generalizability of the findings. Secondly, the fact that the text created within the scope of the study is related to the field of social sciences may have increased the bias in the measurement. For this reason, repeating the measurement on a different text related to the quantitative field in which tables and graphics are used can serve the purpose of testing the conceptual framework used in developing the scale. Third, in this study, predictive and discriminant validity studies of Pamukkale Critical Thinking Skills Scale were not conducted. New studies to be carried out in this context may contribute to the strengthening of the psychometric properties of the scale. Fourth, the structure of the scale was not tested in groups with different characteristics in this study. Future studies, with new research on the measurement invariance of the scale; It can serve the purpose of testing the structure of the scale in groups with different characteristics such as gender, socio-economic level, verbal-numerical domain.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Pamukkale University/Social and Humanities Ethics Committee, 93803232-622.02.

Authorship Contribution Statement

Erdinc Duru: Conceptualization, research design, supervision, fundings, interpretation, literature review, writing, critical review. **Sevgi Ozgungor**: Conceptualization, research design, supervision, fundings, data collection and processing, interpretation, literature review, writing, critical review **Ozen Yildirim**: Research design, supervision, fundings, data collection and processing, data analysis and interpretation, literature review, writing, critical review **Asuman Duatepe-Paksu**: Research design, supervision, fundings, data collection and processing, critical review **Sibel Duru**: Research design, supervision, fundings, data collection and processing, critical review.

Orcid

Erdinc Duru https://orcid.org/0000-0001-7027-4937
Sevgi Ozgungor https://orcid.org/0000-0003-4954-1572
Ozen Yildirim https://orcid.org/0000-0003-2098-285X
Asuman Duatepe-Paksu https://orcid.org/0000-0003-2504-6294
Sibel Duru https://orcid.org/0000-0002-8152-8610

REFERENCES

Abrami, P.C., Bernard, R.M., Borokhovski, E., Wade, A., Surkes, M. ., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*(4), 1102-1134.

Anderson, L.W., & Krathwohl, D.R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives: Complete Edition*. Longman.

Atay, S., Ekim, E., Gökkaya, S., & Sağım, E. (2009). Sağlık Yüksekokulu öğrencilerinin eleştirel düşünme düzeyleri. [Critical thinking tendencies of Health School students] *Sağlık Bilimleri Fakültesi Hemşirelik Dergisi,* 39-46.

Ayberk, B., & Çelik, M. (2007). Watson-Glaser Eleştirel, Akıl Yürütme Gücü Ölçeği'nin (W-GEAYGÖ) üniversite ikinci, üçüncü ve dördüncü sınıf İngilizce bölümü öğretmen adayları üzerindeki güvenlik çalışması. [Reliability study related to the power of Watson-Glaser critical thinking appraisal scale on university second, third and fourth-grade English department teacher candıdates] *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 16* (1), 101-112

Aybek, E.C. (2021). CatIRT tools: A "Shiny" application for item response theory calibration and computerized adaptive testing simulation. *Journal of Applied Testing Technology, 22*(1), 23-27.

Baker, F.B. (2001). The basics of item response theory. College Park, ERIC, Clearinghouse on Assessment and Evaluation.

Batur, Z., & Özcan, H.Z. (2020). Eleştirel düşünme üzerine yazılan lisansüstü tezlerinin bibliyometrik analizi. [Bibliometric analysis of graduate theses written on critical thinking] *Uluslararası Türkçe Edebiyat Kültür Eğitim Dergisi, 9*(2), 834-854.

Bailey, R., & Mentz, E. (2015). IT teachers' experience of teaching–learning strategies to promote critical thinking. *Issues in Informing Science and Information Technology, 12*(1), 141–152.

Bensley, D.A., Crowe, D.S., Bernhardt, P., Buckner, C., & Allman, A.L. (2010). Teaching and assessing critical thinking skills for argument analysis in psychology. *Teaching of Psychology, 37*(2), 91–96. https://doi.org/10.1080/00986281003626656

Bennett, D.A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464–469.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bilican, S.D. (2021). Başarı testlerinin geliştirilmesi ve madde yazımı [Development of achievement tests and item writing] Yıldırım, Ö. ve Kartal, S.K. (Ed.), *Eğitimde Ölçme ve Değerlendirme* [*Measurement and Evaluation in Education]* (125-163) 1. Baskı, Lisans Yayıncılık.

Browne, N., & Freeman, K. (2000). Distinguishing features of critical thinking classrooms. *Teaching in Higher Education. 5*(3), 301-309. https://doi.org/10.1080/713699143

Boyd, E.M., & Fales, A.W. (1983). Reflective learning: Key to learning from experience. *Journal of Humanistic Psychology, 23*(2), 99-117. https://doi.org/10.1177/00221678832 32011

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. K. A. Bollen and J.S. Long (Ed.), *Testing structural equation models* (pp. 136-162). Sage.

Carpendale, J.I., & Lewis, C. (2006). *How children develop social understanding*, Blackwell.

Chalmers, R.P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v 048.i06

Cisneros R .M. (2009). Assessment of critical thinking in pharmacy students. *American Journal of Pharmaceutical Education, 73*(4), 66. https://doi.org/10.5688/aj730466

Cohen, R.J., Swerdlik, M.E., Smith, D.K., & Cohen, R.J. (1992). *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Pub. Co.

Comfort, L.K. (2007). Crisis management in hindsight: cognition, communication, coordination, and control. *Public Administration Review, 67*(1), 189–197.

Crockett, L. (2019). *Future-focused Learning: 10 essential shifts of everyday practice.* Solution Tree Press.

Dewey, J. (1933). *How we think*. DC Herman.

Doğan, N. (2013). Eleştirel düşünmenin ölçülmesi [Measuring the Critical Thinking]. *Cito Eğitim: Kuram ve Uygulama, 22*(1), 29-42.

Doğanay, A., Akbulut-Taş, M., & Erden, Ş. (2007). Üniversite öğrencilerinin bir güncel tartışmalı konu bağlamında eleştirel düşünme becerilerinin değerlendirilmesi. [Assessing university students' critical thinking skills in the context of a current controversial issues]. *Kuram ve Uygulamada Eğitim Yönetimi, 52*(1), 511-546.

Dumitru, D., Bîgu, D., Elen, J., Jiang, L., Railiene, A., Penkauskiene, D., Papathanasiou, I.V., Tsaras, K., Fradelos, E.C., Ahern, A.K., McNally, C., O'Sullivan, J., Verburgh, A.P., Jarošová, E., Lorencová, H., Poce, A., Agrusti, F., Re, M.R., Puig, B., Blanco, P., Mosquera, I., Crujeiras-Pérez, B., Dominguez, C., Cruz, G., Silva, H., & Morais, M.D., Nascimento, M.M., & Payan-Carreira, R. (2018). *A European review on Critical Thinking educational practices in Higher Education Institutions*. http://hdl.handle.net/10197/9865

Duru, E., Duatepe-Paksu, A., Balkıs, M., Duru, S., & Bakay, E. (2020). Examination of 21st century competencies and skills of graduates from the perspective of sector representatives and academicians. *Journal of Qualitative Research in Education, 8*(4), 1059-1079. https://doi.org/10.14689/issn.2148-2624.8c.4s.1m

Dwyer, C.P., Hogan, M.J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity, 12*(1), 43-52. https://doi.org/10.1016/j.ts c.2013.12.004

Ebel, R.L. (1972). *Essentials of educational measurement.* Prentice-Hall.
Eğmir, E., & Ocak, G. (2016). Eleştirel düşünme becerisini ölçmeye yönelik bir başarı testi geliştirme. [Developing an achievement test towards evaluating critical thinking skill]. *Turkish Studies, 11*(19), 337-360.
Ennis, R. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy, 14*(1), 15-24. http://dx.doi.org/10.5840/teachphil19911412
Ennis, R. H., & Millman, J. (1985). *Cornell Critical Thinking Test (Level X)*. Critical Thinking Press & Software.
Facione, P.A. (1990a*). Critical thinking: A statement of expert consensus for purposes of educational assessment and ınstruction. executive summary: "The Delphi Report"*. The California Academic Press.
Facione, P.A. (1990b). *The California Critical Thinking Skills Test-College Level Technical Report 1: Experimental Validation and Content Validity.* The California Academic Press.
Facione, P.A. (1990c). *The California Critical Thinking Skills Test-College Level Technical Report 2: Factors Predictive of Critical Thinking Skills*. The California Academic Press.
Facione, N.C. (1997) *Critical thinking assessment in nursing education programs An aggregate data analysis.* The California Academic Press.
Facione, P.A. (2015). Critical thinking: What it is and why it counts. http://www.student.uwa.edu.au/__data/assets/pdf_file/0003/1922502/Critical-Thinking-What-it-is-and-why-it-counts.pdf
Facione, P.A., & Facione N.C. (1992). *The california critical thinking dispositions inventory.* The California Academic Press.
Fayers, P., & Machin, D. (1995). Factor analysis for assessing validity. *Quality of Life Research*, *4*(5), 424.
Flores, K., Matkin, G.S., Burbach, M.E., Quinn, C.E., & Harding, H. (2012). Deficient critical thinking skills among college graduates: ımplications for leadership. *Educational Philosophy and Theory, 44*(2), 212-230. https://doi.org/10.1111/j.1469-5812.2010.00672.x
Güçlü, G., & Evcili, F. (2021). Sağlık hizmetleri meslek yüksekokulu öğrencilerinin eleştirel düşünme yetileri ve boyun eğici davranış eğilimlerinin incelenmesi. [Investigation of critical thinking qualifications and submissive behavior tendency of health services vocational school students]. *Turkish Journal of Science and Health. 2*(1), 31-39.
Haladyna, T.M. (1997). *Writing test ıtem to evaluate higher order thinking*. Allyn & Bacon.
Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory*. Kluwer-Nijhoff Publishing.
Huber, C.R., & Kuncel, N.R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research, 86*(2), 431-468. https://doi.org/10.3102/0034654315605917
Jacobs, S.S. (1995). Technical characteristics and some correlates of the California Critical Thinking Skills Test, Forms A and B. *Research in Higher Education, 36*(1), 89-108. https://doi.org/10.1007/BF0220776
Irani, T., Rudd, R., Gallo, M., Ricke, J., Friedel, C., & Rhoades, E. (2007) *Critical thinking ınstrumentation manual*. University of Florida.
Lipman, M. (1988) Critical thinking—What can it be? *Educational Leadership, 461*(1), 38–43.
Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, 10(1), 17-29.
Marin, L., & Halpern, D. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity, 6*(1), 1–13.

Mazer, J.P., Hunt, S.K., & Kuznekoff, J.H. (2007). Revising general education: Assessing a critical thinking ınstructional model in the basic communication course. *The Journal of General Education 56*(3), 173-199. https://doi.org/10.1353/jge.0.0000

Mpofu, N., & Maphalala, M.C. (2017). Fostering critical thinking in initial teacher education curriculums: A comprehensive literature review. *Gender and Behaviour, 15*(2), 9342–9351.

Msila, V. (2014). Critical Thinking in open and distance learning programmes: Lessons from the University of South Africa's NPDE Programme. *Journal of Social Sciences, 38*(1), 33–42

Nalçacı, A., Meral, E., & Şahin, İ.F. (2016). Sosyal bilgiler öğretmen adaylarının eleştirel düşünme ile medya okuryazarlıkları arasındaki ilişki [Correlation between critical thinking and media literacy of social sciences pre-service teachers]. *Doğu Coğrafya Dergisi, 21*(36), 1-12. https://doi.org/10.17295/dcd.99051

Norris, S.P., & Ennis, R.H. (1990). *The practitioners' guide to teaching thinking series. Evaluating Critical Thinking*. Hawker Bronlow Education.

Özmen, K.S. (2008). İngilizce öğretmeni eğitiminde eleştirel düşünce: Bir vaka çalışması. [Critical Thinking in English teacher education: A case study]. *Ekev Akademi Dergisi, 12*(36), 253-266.

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), 132-144. https://dx.doi.org/10.52380/ijpes.2022.9.1.561

Parkhurst, H.B. (1999). Confusion, lack of consensus, and the definition of creativity as a construct. *Journal of Creative Behavior, 33*(1), 1–21.

Paul, R. (2005) The state of critical thinking today*, New Directions for Community Colleges, 130*(1), 27–38.

Paul, R., & Elder, L. (2001) Critical thinking: Inert information, activated ignorance, and activated knowledge, *Journal of Developmental Education, 25*(2), 36–37.

Paul, R.W., & Elder, L. (2002). *Critical thinking: Tools for taking charge of your professional and personal life*. Pearson Education Inc.

Paul, R., & Nosich, G. (1991). *A proposal for the national assessment of higher-order thinking.* Paper commissioned by the U.S. Department of Education Office of Educational Research and Improving National Center for Education Statistics.

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

Pascarella, E.T., & Terenzini, P.T. (2005). *How college affects students: A third decade of research*. Jossey-Bass.

Portney, L.G., & Watkins, M.P. (2000) *Foundations of clinical research: Applications to practice.* 2nd Edition, Prentice Hall.

Puig, B., Blanco-Anaya, P., Bargiela, I.M., & Crujeiras-Pérez, B. (2019). A systematic review on critical thinking intervention studies in higher education across professional fields. *Studies in Higher Education, 44*(5), 860-869.

R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Rezaee, M., Farahian, M., & Ahmadi, A. (2012). Critical thinking in higher education: Unfulfilled expectations. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *3*(2), 64-73.

Ruggiero V.R. (1990) *Beyond feelings. A guide to critical thinking*, (3rd ed.) Mayfield Publishing.

Røe, C., Damsgård, E., Fors, T., & Anke, A. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic

musculoskeletal pain. *BMC Musculoskelet Disord, 15*(1), 95. https://dx.doi.org/10.1186/1471-2474-15-95, PubMed 24646065

Sahool, S., & Mohammed, C.A. (2018). Fostering critical thinking and collaborative learning skills among medical students through a research protocol writing activity in the curriculum. *Korean J Med Educ, 30*(2), 109-118. https://doi.org/10.3946/kjme.2018.86

Schafer, J.L. (1999). Multiple imputation: a primer. *Stat Methods in Med.,* 8(1), 3–15. https://doi.org/10.1191/096228099671525676

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

Siegel, H. (1988). *Educating for reason: Rationality, critical thinking, and education*. Routledge.

Snyder, L.G., & Snyder, M.J. (2008). Teaching critical thinking and problem solving skills. *The Journal of Research in Business Education, 50* (2), 90–99.

Stevens, J. (1992). Applied multivariate statistics for the social sciences. Second Edition, Lawrence Erlbaum Associates.

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed.), Allyn and Bacon.

Tolman, E.C. (1932). *Purposive behavior in animals and men*. Century/Random House UK.

Trilling, B., & Fadel, C. (2009). *21st-century skills: Learning for life in our times*. John Wiley & Sons.

Uzuntiryaki- Kondakçı, E., & Çapa-Aydın, Y. (2013). Predicting critical thinking skills of university students through metacognitive self-regulation skills and chemistry self-efficacy. *Educational Sciences: Theory & Practice, 13*(1), 666-670.

Van Laar, E., van Deursen, A.J.A. M., van Dijk, J.A.G.M., & de Haan, J. (2019). Determinants of 21st-century digital skills: A large-scale survey among working professionals. *Computers in Human Behavior, 100*, 93–104. https://doi.org/10.1016/j.chb.2019.06.017

Voogt, J., & Roblin, N.P. (2012). A comparative analysis of international frameworks for 21st-century competencies: implications for national curriculum policies. *Journal of Curriculum Studies, 44*(3), 299–321.

Watson, G., & Glaser, E.M. (1980). *Watson-Glaser critical thinking appraisal*. Psychological Corporation

Williams, R.L., Oliver, R., Allin, J.L., Winn, B., & Booher, C.S. (2003). Psychological critical thinking as a course predictor and outcome variable. *Teaching of Psychology, 30*(3), 220–223. https://doi.org/10.1207/S15328023TOP3003_04

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(1), 187-213.

APPENDIX

Sample Multiple Choice Questions and Some Options

1. Aşağıdakilerden hangisi metnin amacını **en iyi** açıklar?

Metnin amacı,

a) Aşıların otizme neden olup olmadığını göstermektir.
b) Aşı ile otizm arasındaki ilişkiye dair bazı tartışmaları sunmaktır.
c) Otizmin nedenlerini yapılan araştırmaları karşılaştırarak açıklamaktır.

2. Hangi seçenekte metinden çıkarılabilecek gerekçelerin tamamı birlikte verilmiştir?

a) Son 20 yılda gelişen teknolojiyle birlikte otizm vakalarının artması- Aşı tartışmaları sonucunda ailelerin çocuklarına aşı yaptırmaması- Aşıların içindeki cıvanın otizme neden olması- Fazla miktarda balık tüketimi olması
b) Otizmli 12 çocukla yapılan araştırma sonuçları- Gelişen teknolojinin insan sağlığını tehdit etmesi- Cıva içeren balıkların tüketiminin nörolojik hastalıklara neden olması- Otizm ile çocuklardaki alüminyum oranı arasındaki ilişki
c) Otizmli çocukların çoğunluğunun aşılı olması- Otizm ve aşı arasındaki ilişkilerin araştırma sonuçlarına dayanması- Son yıllarda otizmin artması- Sayısal verilerle otizm ile aşı arasındaki ilişkinin desteklenmesi

3. Aşağıdakilerden hangisinde bebeklerine aşı yaptırma konusunda kararsız kalan ebeveynlere, metinden çıkarılacak gerekçelere dayalı **en uygun** öneri verilmiştir?

a) Araştırma sonuçlarından çıkarılacağı gibi aşı yaptırmamalarını öneririm. Çünkü aşı olmayan birçok insan günümüzde sağlıklı bir şekilde hayatlarına devam edebilmektedir.
b) Farklı kaynaklardan araştırmalarını ve uzmanlara sormalarını öneririm. Çünkü aşı yaptırırlarsa otizm olma, yaptırmazlarsa bulaşıcı hastalıklara yakalanma olasılığı söz konusudur.
c) Farklı kaynaklardan araştırıp, uzmanlara danışmalarını, sonucunda aşı yaptırmalarını öneririm. Çünkü aşı yapılmadığı takdirde bulaşıcı hastalıklarda artış gözlenmiştir.

4. Aşı yaptırmayı savunan bir çocuk doktorunun bu metni okuduktan sonraki düşüncelerini aşağıdakilerden hangisi **en iyi** yansıtır?

a) Aşılar gereklidir. Ancak aşıların olası yan etkileri ve ebeveynlerin kaygıları dikkate alındığında başka araştırmaların da incelenmesi önemlidir.
b) Aşı önemlidir, aşı olmayan çocukların bulaşıcı hastalıklara karşı bağışıklıkları düşük olduğundan, bebeklere küçük yaştan itibaren aşı yapılmalıdır.
c) Çocukların daha sağlıklı büyüyebilmesi için bazı aşılar zamanında yapılmalıdır ve en kısa sürede tekli aşı sistemine geçilmelidir.