International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 690–714
https://doi.org/10.21449/ijate.744640
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Assessment of the Fifth-Grade Students' Science Critical Thinking Skills through Design-Based STEM Education

Ayse Savran Gencer [1,*], **Hilmi Dogan** [2]

[1]Department of Science education, Faculty of Education, Pamukkale University, Denizli, Turkey.
[2]Ministry of National Education, Antalya, Turkey.

ARTICLE HISTORY
Received: May 28, 2020
Revised: Sept. 10, 2020
Accepted: Oct. 23, 2020

KEYWORDS
STEM education,
Science critical thinking,
Design-based,
Living things,
Force and friction

Abstract: Critical thinking has been one of the 21st-century skills consistently associated with students' future career advancement as a positive student outcome of STEM education. The aim of the study is to develop and validate science critical thinking skill instruments to assess the improvement in the subject of living organisms and force and friction through design-based STEM education. In this design-based research study, the student's modules were developed by the integrated STEM education principles involving the activities and worksheets in line with the frame of critical thinking approach. The kappa statistics for content validity, exploratory and confirmatory factor analysis for construct validity, and item and reliability analysis for the quality of items were used in the development stage of instruments. The results of these analyses endorsed the 15 two-tier item for each test of Living Things Critical Thinking (LTCT) and Measuring Force and Friction Critical Thinking (MFFCT) as unidimensional constructs to produce valid and reliable data to measure the fifth grade students' critical thinking skills in the related science content. Comparing the pre and post applications of instruments in the study group indicated that STEM modules improved the students' science critical thinking skills such as interpretation, analysis, and inference. In this respect, developing and validating instruments to assess the integrated critical thinking skills will contribute to the empirical examination of this construct within the context of school science learning.

1. INTRODUCTION

Rapid changes in the flow of knowledge in today's world have led to nations to revise science education programs and science teaching goals in such a way of cultivating individuals who are able to produce knowledge and use it functionally in their lives by contributing to society and culture with the skills of problem-solving, critical thinking, entrepreneurship, decision making, collaboration, communication, and empathy (e.g., Ministry of National Education [MoNE], 2018). To accomplish these goals, the initiatives of integrating science, technology, engineering, and mathematics (STEM) have been appropriated as interdisciplinary approach by involving learning about knowledge, skills, beliefs and values from more than one STEM discipline through the collaborative efforts of students and teachers (Baharin, Kamarudin, & Manaf, 2018; Çorlu, Capraro, & Capraro, 2014; Ergün & Külekci, 2019; Öner et al., 2014;

CONTACT: Ayşe Savran Gencer ✉ asavran@pau.edu.tr Department of Science education, Faculty of Education, Pamukkale University, Denizli, Turkey.

ISSN-e: 2148-7456 /© IJATE 2020

Wang, Moore, Roehning, & Park, 2011). In particular, STEM teaching can be more meaningful when embedded in real-life problems with challenges in a manner of integrity for extending students' motivation and persistence to learn and succeed in science (Honey, Pearson, & Schweingruber, 2014).

Critical thinking has been one of the 21st-century skills consistently associated with students' future career advancement as a positive student outcome of STEM education (Next Generation Science Standards [NGSS] Lead States, 2013). A great deal of literature from many countries provide insight on how design based STEM learning activities engage students to solve real-world problems through investigating and collaborating with their peers in establishing an effective learning environment to foster critical thinking skills (Baharin et al., 2018; Duran & Şendağ, 2012; Mutakinati, Anwari, & Yoshisuke, 2018; Oonsim & Chanprasert, 2017; Rahmawati, Ridwan, Hadinugrahaningsih, & Soeprijanto, 2019; Waddell, 2019). For instance, a study with Japanese middle school students by using STEM education through project-based learning to solve the need for clean water in the future reported that students' overall critical thinking skills developed up to the category of the average thinker (Mutakinati et al., 2018). In the Indonesia context, Rahmawati et al. (2019) explored that integrating STEAM approach into chemistry learning within real-life problems provided opportunities for students to improve their critical thinking skills. In Thailand, Oonsim & Chanprasert (2017) indicated an average increase of critical thinking skills by using STEM education in the subject of physics for secondary school students. For the United States, Duran & Şendağ (2012) reported a significant effect of STEM experiences enhanced with information technology on the improvement of urban high school students' critical thinking.

1.1. The Theoretical Framework

Critical thinking is the process of mentally acting on something by "making reasoned judgments" (Beyer, 1995, p.8). Facione (1990) defines critical thinking as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" (p. 2). Thinking skills in education settings generally involve activities of comparing and contrasting, classifying, predicting, generating original ideas, cause and effect, decision making, uncovering assumptions, and determining the reliability of sources of information (Swartz, Costa, Beyer, Reagan, & Kallick, 2008). Critical thinking includes process skills of "analysing, evaluating, or synthesizing relevant information to form an argument or reach a conclusion supported with evidence" (Reynders, Lantz, Ruder, Stanford, & Cole, 2020, p.4). Critical thinking enables individuals to develop their way of thinking about any subject, content, or problem by skilfully handling thought-specific structures and assigning intellectual standards to them. Then, these individuals can use the principles that help them to improve their thinking while analysing and evaluating the problems or their thoughts (Gencer & Boran, 2017). In addition, critical thinking skills are required in the process of analysing possible solutions during the problem solving, evaluating the consistency between alternatives during decision making or predicting the results of the decision (Dilekli, 2019).

Beyond its wide range of definitions, there has been a dichotomy in the construction of critical thinking as domain-general versus domain-specific (Ennis 1989, Facione 1990; National Research Council [NRC], 2011; Swartz et al., 2008; Willingham, 2008). According to the report by NRC (2011), the predominant view on domain-specific construct advocates that critical thinking coevolves with the increasing content knowledge and cannot be transferred spontaneously from one subject matter to another. Willingham (2008) ascertains specific types of critical thinking to the extent to which they are characterized by different subject matters. Bailin (2002) points to the contextual nature of critical thinking in science education due to the fact that focusing on the concepts, tasks, problems, and issues in the science curriculum initiate

critical thinking by collecting knowledge from observation, classification, correlation, causation, hypothesis, inference and prediction as well as background knowledge of students for critical analysis, interpretation, and evaluation. Students cannot learn spontaneously how to think in each subject matter and therefore the ways should be modelled with the characteristic of the different subject matter by giving opportunities to practice in the context of the related classroom tasks (Willingham, 2008).

Such specific types of skilful thinking and mental behaviours need to be taught students explicitly and by direct instruction to be effective thinkers (Swartz et al., 2008). Regarding the instruction of critical thinking skills, Ennis (1989) classifies approaches into four types. The general approach involves teaching critical thinking skills in a separate course without a specific subject matter. According to the infusion approach, students are involved in the explicit teaching of critical thinking skills process in a specific subject matter. In the immersion approach, a subject course is organised to teach critical thinking, but critical thinking principles are not given explicitly. The mixed model approach combines a general approach with infusion or immersion approach. In this research, we have adopted the infusion approach to teach critical thinking skills in science-domain. The student's modules for this study were designed by the integrated STEM education principles involving activities and worksheets in reference to Swartz et al.'s (2008) frame of critical thinking approach.

1.2. The Significance and Purpose of the Study

A limited research has been recently conducted to develop and assess critical thinking skills of students in the specific content knowledge such as mathematics (Harjo, Kartowagiran, & Mahmudi, 2019; Kuş & Çakıroğlu, 2020), chemistry (Reynders et al., 2020; Sadhu & Laksono, 2018), physics (Asysyifa, Jumadi, Wilujeng, & Kuswanto, 2019; Mabruroh & Suhandi, 2017), and science (Mapeala & Siew, 2015; Sya'bandari, Firman, & Rusyati, 2017). At primary level, such a study by Mapeala and Siew (2015) developed a science critical thinking test to measure the critical thinking skills of the fifth-grade students in the theme of physical sciences. At secondary school level, Sya'bandari et al. (2017) constructed a science virtual test to measure the seventh-grade students' critical thinking in the matter and heat topic. For high school students, there were integrated assessment instruments to measure critical thinking skills in the concepts of chemical equilibrium (Sadhu & Laksono, 2018) and sound waves (Mabruroh & Suhandi, 2017). In reaction to the shortage of subject-specific construct of critical thinking skills in STEM fields, this study will contribute the broadening the scope of science subjects to be taught for integrated critical thinking skills.

Another issue associated with teaching critical thinking skills is the importance of developing students' critical thinking skills at an early age. In essence, critical thinking skills should be embedded in the science curriculum from beginning in the early grades of schooling and growing in complexity and sophistication throughout the grades (Bailin, 2002; Wicaksana, Widoretno, & Dwiastuti, 2020). In doing so, the current study can contribute to the development and assessment of the early grade students' science critical thinking skills in informing science educators and teachers about how to design an effective learning environment to teach science critical thinking skills in their classroom. Due to the fact that the importance of critical thinking skill has been appreciated as one of the higher-order thinking skills to be assessed in international exams (e.g., Programme for International Student Assessment [PISA]) and national exams in Turkey (e.g., High School Pass Exam), further investigations need to be done to construct and measure more accurately integrated critical thinking skills in science learning. In an effort to attain these goals, the present study aims to develop and validate science critical thinking skill instruments to assess the improvement in the subject of living organisms and force through design-based STEM education.

2. METHOD

The current research is a part of a larger dissertation study based on design-based research consisting of the preliminary research phase, the prototyping phase (the iterative design phase), and the assessment/reflective phase as proposed by Abdallah & Wegerif (2014). Based on the preliminary phase, STEM modules were developed to provide students the learning opportunities to explore both engineering design principles and learning outcomes in the units of "Living Things World" and "Measuring Force and Friction". Worksheets in the related STEM modules were constructed in line with the critical thinking frame of Swartz et al. (2008) including analysing ideas in terms of comparing and contrasting, classifying, sequencing, determining parts/whole relationship, identifying causal relationship and further analysing arguments, drawing a conclusion, making a decision and problem solving to integrate learning tasks with critical thinking skills. Appendix I indicates examples of worksheets to integrate specified critical thinking skills with learning tasks in the modules.

The iteration cycles include the eight-step engineering design process of the Massachusetts Department of Education (2006). The engineering design cycle consists of identifying the need or problem, researching the need or problem, developing a possible solution, selecting the best possible solution, constructing a prototype, testing and evaluating the solution, communicating the solution, and redesigning. The study was conducted for sixty-one hours during the science lessons in the first term of the school year of 2018-2019 by the second author of this research.

The students' learning modules which were designed with STEM education approach applied to the study group as an intervention. The assessment/reflective phase involved an assessment of the instruments as a pre- and post-test to collect data about the effectiveness of the STEM modules on the student's integrated critical thinking skills.

2.1. Study Group

In the first stage of the preparation phase, the sample for the pilot study was needed for developing the science domain instruments. The data were obtained from the sixth grade students of (N = 147) for Living Things Critical Thinking (LTCT)-Test and (N = 116) for Measuring Force and Friction Critical Thinking (MFFCT)-Test studying in three different public secondary schools located in Antalya.

In the second phase of the prototyping phase, the study group was chosen with a convenience sampling method (Patton, 2014) by considering the school where the second researcher worked as a teacher. The study was carried out with 22 students (10 girls, 12 boys) who were 10-11 years old attending the fifth grade at the public secondary school in Antalya, Turkey.

2.2. The Instruments

The instruments were developed in order to evaluate the integrated critical thinking skills within school science learning through design-based STEM education. LTCT and MFFCT tests focused on the three elements of the critical thinking skills including interpretation, analysis, and inference as proposed by Facione (1990) within the contents of the current science education curriculum in Turkey (MoNE, 2018). Each of the final version of instruments consists of 15 two-tier multiple-choice items. The first-tier item includes content questions with four choices and the second tier includes a blank for the first part to allow students to explain the reason why they choose the option in the first tier (Griffard & Wandersee, 2001). The items of open-ended two-tier multiple-choice tests were scored as 3 (the right answer -the right reason), 2 (the right answer- partly correct reason), 1 (the right answer- the wrong reason), 2 (the wrong answer-the right reason), 1 (the wrong answer- partly correct reason), and 0 (the false answer-the wrong reason). In this study, one point is given for the students who can write a partially correct reason despite their wrong answers in addition to the commonly used scoring in the

literature. A guideline was prepared for students and practitioners regarding the duration of the test, scoring method and what they are expected to explain in the second tier.

2.2.1. *The Instrument Development Process*

In the initial versions, 19 two-tier items for LTCT-Test and 20 two-tier items for MFFCT-Test with four options were developed. Some of the cognitive critical thinking skills and sub-skills compiled by Facione's (1990) Delphi report and learning objectives in the science curriculum (MoNE, 2018) in Turkey were taken into consideration as a guide for the development of the science-domain critical thinking tests. In the unit of Living Things World, students are expected to give examples of living things and classify them according to their similarities and differences as microscopic organisms, fungi, plants, and animals. In the unit of Measuring Force and Friction, students are expected to measure the magnitude of the force with a dynamometer, give examples of friction force from daily life, discover the effect of friction force on motion in various environments, do experiments about the effect of friction force on motion on rough and slippery surfaces, and generate new ideas to increase or decrease friction in everyday life (MoNE, 2018).

LTCT-Test contains critical thinking constructs of interpretation (4 items), analysis (4 items), and inference (7 items). Table 1 indicates the core and sub-skills of the critical thinking constructs within the science content for LTCT-Test. MFFCT-Test contains critical thinking constructs of interpretation (5 items), analysis (3 items), and inference (7 items). Table 2 indicates the core and sub-skills of the critical thinking constructs within the science content for MFFCT-Test. A sample item was given in Figure 1 and Figure 2 for LTCT-Test and MFFCT-Test, respectively.

Table 1. *The integrated critical thinking skills and the science content for LTCT-Test*

Item	Core Skills of Critical Thinking	Sub-skills	Scientific content
1-3-7-10	Interpretation	Categorization	Identifying the distinctive and/or common features of living things from image/diagram/text to classify them.
2	Analysis	Examining ideas	Comparing and contrasting living things by classifying them in terms of criteria and revealing the relations using the data presented in the graph.
8		Analysing arguments	Recognizing the difficulties in classifying living things and distinguishing the rejecting or supporting reasons for the claims regarding the classification.
11-14			Distinguishing the rationale for rejecting or supporting the claim regarding the classification of living things.
4-12-15	Inference	Drawing conclusion	Drawing a conclusion about the function of the structure by observing the structure of living things.
5			Drawing a conclusion that scientific knowledge is tentative by using relevant information/data
6			Drawing a conclusion about how scientific knowledge is formed and the process through which knowledge is passed.
13		Querying evidence	Deciding the accuracy of classification of living things based on evidence.
9		Conjecturing alternatives	Identifying the hypothesis tested by obtaining the variables that affect the growth of bacteria from a given experimental setup.

Item 11. **Core skill: Analysis** **Sub-skill: Analysing arguments**

Melek and Cemre, who have visited the zoo, come to the section with seals. Melek carefully examines this creature because she has seen those animals for the first time and says to Cemre:

Melek: It can breathe on land. I think this is a mammal.
Cemre: Frogs can breathe on the land too, but they are not a mammal. It has got fins and swims like a fish. I think that it must be a fish.
Melek: But it has not got scales. Instead of them, it seems to have short and stiff hairs

Melek and Cemre cannot decide whether the seal is a mammal or fish. They then decide to ask a zoologist working at the zoo. After the zoologist gives them information, Melek and Cemre are convinced the seal is a mammal.
According to the text, what might the zoologists have told Melek and Cemre?
A. Seals feed on other fishes
B. Seals have skeletons
C. Seals feed their offsprings with milk
D. Seals do not have gills
How did you decide that the option you marked was correct? Please explain.
.....

Figure 1. *A sample item for an open-ended two-tier multiple-choice question in LTCT-Test*

Item 14. **Core skill: Inference** **Sub-skill: Drawing conclusion**

Predator birds can see their prey while flying high thanks to their sharp eyes. While they are searching for prey, they spread their wings as wide as they can, and they do not have to flap them for a long time. As soon as they see their prey, they close their wings slightly and dive to catch the prey.

In the images below, Figure I shows a predator bird searching its prey, and Figure II shows the bird diving to catch its prey.

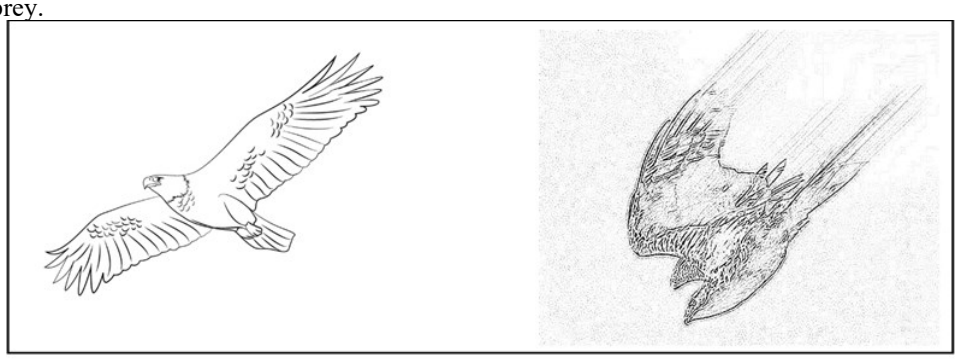


Figure I Figure II

Which of the following can be reached for the two given conditions accordingly?
A. The air resistance affecting the predator bird is greater when its wings are wide open.
B. The bodyweight of the predator bird decreases while it flies with open wings.
C. The gravity affecting the predator bird increases when the bird closes its wings.
D. The air resistance exerted on the predator bird increases when the bird closes its wings.

How did you decide that the option you marked was correct? Please explain.

Figure 2. *A sample item for an open-ended two-tier multiple-choice question in MFFCT-Test*

Table 2. *The integrated critical thinking skills and the science content for MFFCT-Test*

Item	Core Skills of Critical Thinking	Sub-skills	Scientific content
2	Interpretation	Clarifying meaning	Obtaining the magnitude of the force acting on the objects from the given images, display the data with a graph.
12		Categorization	Explaining by obtaining the magnitude of the force acting on the objects from the given graph.
9-13			Classify the applications that increase or decrease the friction force in daily life according to their similar and different features.
5	Analysis	Analysing arguments	Distinguishing the justifications for rejecting or supporting the argument regarding the relation between air friction with the surface area.
8-15			Distinguishing the rationale for rejecting or supporting the argument about the effects of friction force in daily life.
10			Determination the interrelation between the supplied parts and each other in an experiment on the relation between the spring thickness and the sum of its extension
1	Inference	Conjecturing alternatives	Obtaining the magnitude of the forces from the given evidence, making inference by comparing the data.
3		Drawing conclusion	Drawing a conclusion by using empirical data confirming or falsifying the claims regarding the measurement of the magnitude of the force.
4			Obtaining the data about the magnitude of the forces acting on the objects from the visuals, comparing weights and determinate the relationship between the magnitudes.
6			Drawing a conclusion about the tested hypothesis from the result of the experiment that friction force depends on the type of surface/surface area.
7			Identifying and distinguish the causes that help to support the outcome of friction-induced events in everyday life.
11			Drawing a conclusion by using empirical data confirming or falsifying the claims regarding the factors affecting air friction.
14			Identify and distinguish the causes that help support the outcome of friction-induced events in everyday life.

2. 3. Data Analysis

While the instruments were being developed, validity, reliability and item analysis were conducted in order to obtain information for each item whether to use, revise or eliminate the faulty items (Whiston, 2012). For content validation, the written items were examined by a panel of experts (n = 6) consisting of two science teachers, one academician in the field of science education, two academicians in curriculum specialized in critical thinking and one academician in the field of measurement and evaluation. Kappa statistics were used to assess the opinions of experts on the items in terms of the relevance to the content, construct, grade level, and clarity. The kappa statistics were also used to assess the coding of the open-ended parts of two-tier questions.

For construct validity, FACTORv.10.10.01 software was used to determine dimensionality and structure testing with regard to Exploratory Factor Analysis (EFA) carried with optimal implementation of Parallel Analysis (PA) (Timmerman, & Lorenzo-Seva, 2011) based on Polychoric Correlations Matrix (PCM). In order to confirm the unidimensionality of the data, the Confirmatory Factor Analysis (CFA) was applied by using LISRELv.8.80 software.

The obtained data were analysed by TAP 19.1.4 software to carry out item statistics for the first-tier items scored dichotomously (0 and 1) of both LTCT-Test and MFFCT-Test. In this research, the difficulty index (p value) and item discrimination point biserial coefficient (r_{pb}) values were calculated. For the reliability assessment to examine the internal consistency of an

instrument, both Kuder–Richardson formula known as *KR-20* was used to calculate reliability for the first tier of the test items scored as dichotomous (0 and 1) and Cronbach's alpha coefficient was used to calculate for the integrated assessment of items with open-ended second tier scored as polytomous (0, 1, 2, and 3).

After the tests were applied on the study group of this research as a pre- and post-test in order to determine the impact of the intervention, the collected data were analysed by Wilcoxon Test by using SPSS v.22 software programme.

3. RESULT / FINDINGS

In this section, the results of kappa statistics, item statistical analysis, exploratory and confirmatory factor analysis, and Wilcoxon test are presented.

3.1. Content Validity of the Instruments

In the preliminary version of the instruments, 19 two-tier items for LTCT-Test and 20 two-tier items for MFFCT-Test were validated by a panel of expert judges. Each item of the tests was evaluated by the experts considering a) relevance of the item with the content b) relevance of the item with the critical thinking skills c) clarity of the item d) relevance of the item with the grade level. The three attributes of the items were rated in a three-point Likert scale format (1 = not relevant; 2 = partly relevant; 3 = relevant). Also, a blank section for each item is allocated for experts to comment on each item. The modified kappa statistic was computed to estimate the agreement between the experts indicated beyond the chance on item level content reliability (Polit, Beck & Owen, 2007). The probability of chance agreement (P_c) is first computed with formula 1 and to compute modified kappa statistic (k*) inserted into formula 2.

$$PC = \left[\frac{N!}{N_G\,!\,(N - N_G)!}\right] \cdot \left[\frac{1}{2}\right]^N \quad (1)$$

N: Number of experts
N_G: Number of agreements rated relevant

$$Kappa = \frac{\left(\frac{N_G}{N}\right) - P_C}{1 - P_C} \quad (2)$$

The calculated *PC* and the k^* values for LTCT-Test and MFFCT-Test are displayed in the Table 3 and Table 4, respectively. If the kappa value is between (.60 ≤ kappa ≤ .74), the agreement among experts is good. If the kappa value is greater than .75, the agreement among experts is perfect (Fleiss, 1981, as cited in Yurdugül ve Bayrak, 2012). As regards to these criteria, the modified kappa values of items (Items 4, 9, 15, and 19) which were lower than .60 for LTCT-Test were eliminated from the test. The rest of the items all had modified kappa values which were greater than .75. Consequently, it can be interpreted that the agreement among experts was perfect for these items.

Table 3. *The kappa statistics for content validity of LTCT-Test*

Item	a. Relevance of the item with the content					b. Relevance of the item with the critical thinking skills					c. Clarity of the item					d. Relevance of the item with the grade level				
	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*
	R	PR	NR			R	PR	NR			R	PR	NR			R	PR	NR		
1	6	0	0	0.02	1.00	6	0	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00
2	5	1	0	009	0.82	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
3	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
4	4	0	2	0.23	0.56	3	1	2	0.31	0.27	6	0	0	0.02	1.00	5	1	0	0.09	0.82
5	6	1	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
6	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
7	6	0	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
8	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
9	5	1	0	0.09	0.82	3	1	2	0.31	0.27	5	1	0	0.09	0.82	6	0	0	0.02	1.00
10	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
11	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
12	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
13	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
14	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	6	0	0	0.02	1.00
15	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
16	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
17	5	1	0	0.02	1.00	5	0	1	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
18	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
19	4	0	2	0.23	0.56	4	1	1	0.23	0.56	5	1	0	0.09	0.82	5	1	0	0.09	0.82

R: relevant, PR: partly relevant NR: not relevant, P_c: probability of chance relevant, k^*: modified kappa value

Similarly, the modified kappa values of items (Item 7, 12, 13, 16, and 20) were lower than .60 for MFFCT-Test were eliminated from the test. The rest of the items had modified kappa values which were greater than .75. Consequently, it can be interpreted that the agreement among experts was perfect for these items.

Table 4. *The kappa statistics for content validity of MFFCT-Test*

Item	a. Relevance of the item with the content					b. Relevance of the item with the critical thinking skills					c. Clarity of the item					d. Relevance of the item with the grade level				
	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*
	R	PR	NR			R	PR	NR			R	PR	NR			R	PR	NR		
1	6	0	0	0.02	1.00	5	1	0	0.09	0.82	5	1	0	0.09	0.82	6	0	0	0.02	1.00
2	5	1	0	009	0.82	5	1	0	0.09	0.82	6	0	0	0.02	1.00	5	1	0	0.09	0.82
3	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
4	6	0	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
5	6	1	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
6	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
7	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	6	0	0	0.02	1.00
8	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
9	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
10	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
11	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
12	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	5	1	0	0.09	0.82
13	4	2	0	0.23	0.56	3	1	2	0.31	0.27	5	1	0	0.09	0.82	5	1	0	0.09	0.82
14	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
15	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
16	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	6	0	0	0.02	1.00
17	5	1	0	0.02	1.00	5	0	1	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
18	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
19	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
20	4	0	2	0.23	0.56	3	1	2	0.31	0.27	6	0	0	0.02	1.00	5	1	0	0.09	0.82

R: relevant, **PR**: partly relevant **NR**: not relevant, P_c: probability of chance relevant, k^*: modified kappa value

In addition, kappa statistics were calculated for the answers of the open-ended parts of two-tier questions. For this purpose, the second author of the study scored the students' answers at two different times to provide interrater reliability for the consistency of between two scores by a single rater. The kappa values of .883 and .886 were calculated for LTCT-Test and MFFCT-Test, respectively. The level of kappa coefficient indicated that there is excellent consistency of the scores.

3.2. Item Statistical Analysis

The multiple choice first-tier items were scored as a dichotomous variable (0 and 1), and analysis was carried out with the Test Analysis Program (TAP) (Brooks & Johanson, 2003). Coaley (2010) defined that "the difficulty indicator, known as the *p* value, represents the percentage of participants who have answered an item correctly and is calculated by dividing the number of people getting it right by the total number who attempted it" (p.38). An item difficulty index can range from .00 (meaning no one got the item correct) to 1.00 (meaning everyone got the item correct). Whiston (2012) points out that "item difficulty does not really indicate difficulty; rather, because it provides the proportion of individuals who got the item correct, it shows how easy the item is" (p.71). According to the Coaley (2010), "if all is well, the mean item *p* value is about .50 indicates moderate difficulty level…But it does not mean that a mean *p* value of .50 is always appropriate because a high level assessment of cognitive ability may need more difficult items and, therefore, a lower mean value (is preferable)" (p.38).

The item discrimination analysis indicates that each item of the test is related to the overall test performance (Haladayna, 1999; Nunnally & Bernstein, 1994). The discrimination value can be decided by using the point biserial coefficient (r_{pb}) that compares correct and incorrect answers for each item statistically with overall test score performance (Polit & Hungler, 1999). If the item discrimination value is greater than ($r_{pb} \geq .40$), item is very good or perfect; between ($.30 \leq r_{pb} \leq .39$), it is reasonable good; between ($.20 \leq r_{pb} \leq .29$), it is marginal but acceptable; and lower than ($r_{pb} \leq .19$), it is weak and should not be included in the test (Crocker & Algina, 1986; Ebel & Frisbie, 1991; Wiersma & Jurs, 2005).

The result of the TAP analysis produced the item difficulty index (*p*) and point biserial coefficient (r_{pb}) value to determine the discrimination index of items which are displayed for LTCT-Test and MFFCT-Test in Figure 3 and Figure 4, respectively.

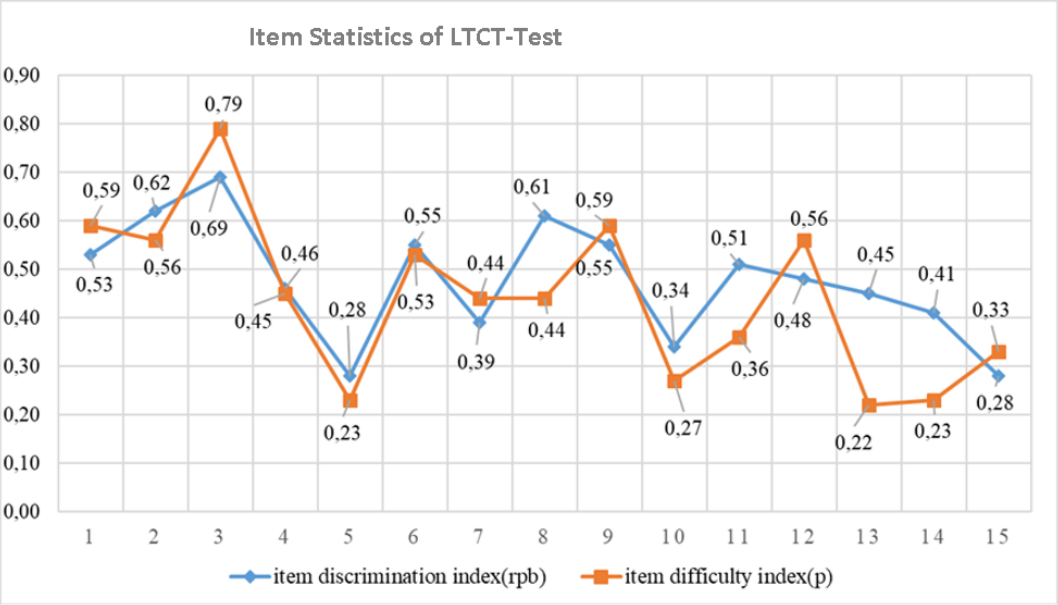


Figure 3. *The graph of item statistics of LTCT-Test*

Figure 3 indicates item difficulty (*p*) values ranging from .22 to .79 for LTCT-Test. The result of the analysis indicates that the test was formed with different difficulty levels of items. The average item difficulty value was calculated as .44 for LTCT-Test. Figure 3 indicating point biserial coefficients for 11 items were greater than .40 and for two items were greater than .30 that these items had perfect and good discrimination values. Item 5 and 15 had a point biserial value of .28, then these two items were decided to conserve in the test, but they should be revised as regards the criteria of ($.20 \leq r_{pb} \leq .29$). Overall, all the items with the value of ($r_{pb} \geq .20$) were determined to be included in LTCT-Test with the average discrimination value of .48.

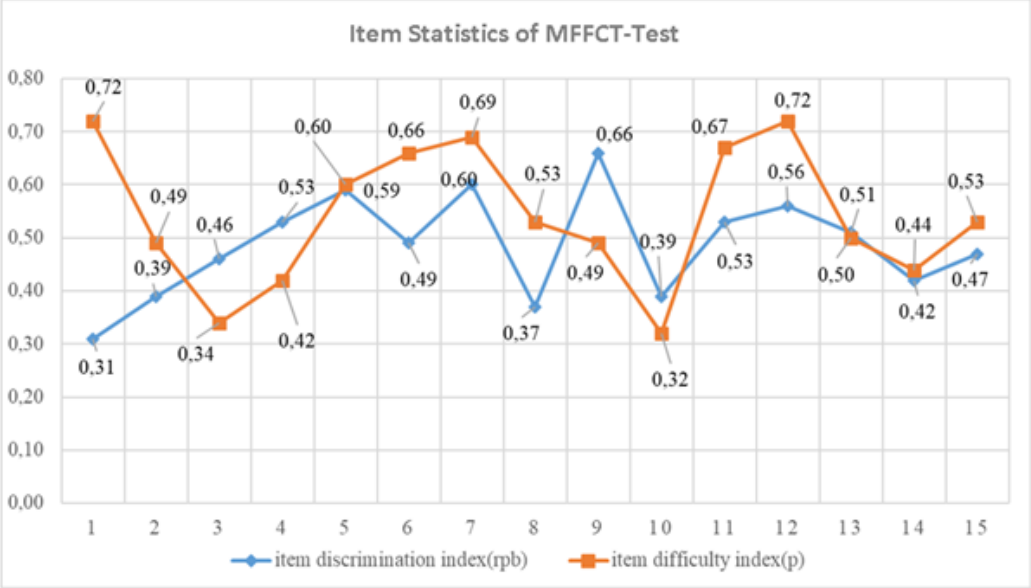


Figure 4. *The graph of item statistics of MFFCT-Test*

Figure 4 indicates item difficulty (*p*) values ranging from .32 to .72 for MFFCT-Test. The average item difficulty value was calculated as .54 for MFFCT-Test. Figure 4 indicating point biserial confidents for 11 items were greater than .40 and for four items were greater than .30 that these items had perfect and good discrimination values. Overall, all the items with the value of ($r_{pb} \geq .20$) were appreciated to stay in MFFCT-Test with the average discrimination value of 0.49.

Reliability analyses of the instruments were tested using both *KR-20* formula and Cronbach's alpha coefficient. The value of reliability coefficient with greater than .70 indicates the test is reliable (Frankel & Wallen, 2008). As regards to this criterion, reliability analysis of *KR-20* for the first tier of the tests produced the acceptable value of .76 and .77 for LTCT-Test and MFFCT-Test, respectively. Also, reliability analysis of Cronbach's alpha coefficient for entire tests as polytomous produced the value of .79 and .91 for LTCT-Test and MFFCT-Test, respectively.

3. 3. Exploratory Factor Analysis of Instruments

Exploratory factor analysis with parallel analysis (PA) based on the polychoric correlations matrix (PCM) was carried out independently for both tests to determine the number of dimensions. Unidimensionality of the tests were determined by the values of Unidimensional Congruence (UniCo > .95), Explained Common Variance (ECV> .85), and Mean of Item Residual Absolute Loadings (MIREAL< .30) for the overall the test as well as for each item (Ferrando & Lorenzo-Seva, 2017). Another evidence for unidimensionality considered in studies is 20% or more explained variance by the first factor with 4 to 5 times greater eigenvalue

when compared to the second factor's eigenvalue (Arıcak, Avcu, Topçu, & Tutlu, 2020; Deng, Wells, & Hambleton, 2008; Hattie, 1985; Yalaki et al., 2019). Prior to factor analysis, Kaiser-Meyer-Olkin (KMO) with the critical value greater than .60 and significant result of Barlett Sphericity test ($p < .05$) were ensured for the convenience of data (Bursal, 2017).

The computed KMO value of .774 with greater than the critical value of .60 and the significant result of Barlett's test of sphericity ($\chi 2(105) = 603.9$, $p = .000010$) for LTCT-Test indicated that the data were appropriate for factor analysis. The result of the parallel analysis at 95% confidence intervals indicated that the values of UniCo = .925 (.918 ≤ UniCo ≤ .948), ECV = .769 (.754 ≤ ECV ≤ .821) and MIREAL= .224 (.199 ≤ MIREAL ≤ .220) provided evidence for the unidimensionality of the test. The parallel analysis suggested a one-factor structure that explained 30.5% of the variance of the test scores with eigenvalue of 4.57. This eigenvalue of the first factor was approximately 3 times the eigenvalue of the second factor that also confirmed the one-factor structure. Büyüköztürk (2012) suggested the cut-off point for factor loadings to be at least .30. The LTCT-Test included 14 items with loadings ranging from .357 to .727 and one item loadings with .258 retained in the test because of the content contribution. As a result of the exploratory factor analysis, the data can be threatened as essentially unidimensional for the LTCT-Test.

The computed KMO value of .834 with greater than the critical value of .60 and significant result of Barlett's test of sphericity ($\chi 2(105) = 999.1$, $p = .000010$) for MFFCT-Test indicated that the data were appropriate for factor analysis. The result of the parallel analysis at 95% confidence intervals indicated that the values of UniCo = .978 (.963 ≤ UniCo ≤ .991), ECV = .884 (.860 ≤ ECV ≤ .930) and MIREAL = .213 (.157 ≤ UniCo ≤ .258) provided the unidimensionality for the test. The parallel analysis suggested one-factor that explained 49.2% of the variance of the test scores with eigenvalue of 7.37. This eigenvalue of the first factor was approximately 6 times the eigenvalue of the second factor that also confirmed the one-factor structure. The MFFCT-Test included 15 items with loadings ranging from .561 to .823. As a result of the exploratory factor analysis, the data can be threatened as essentially unidimensional for MFFCT-Test.

3. 4. Model-Data Fit Analysis of Instruments

Confirmatory factor analysis (CFA) was used to determine whether the existing structure of the instruments confirms the one-factor model (Doğan, 2019; Whiston, 2012). The CFA tests the theory rather than producing a theory (Stevens, 2002). The path analysis was used to confirm the structure of the test as a technique of the Structure Equation Model (SEM) (Awang, 2012; Hair, Black, Babin, & Anderson, 2009). Figure 5 and Figure 6 present the results of the standardized one-factor model solution for the LTCT-Test and MFFCT-Test, respectively.

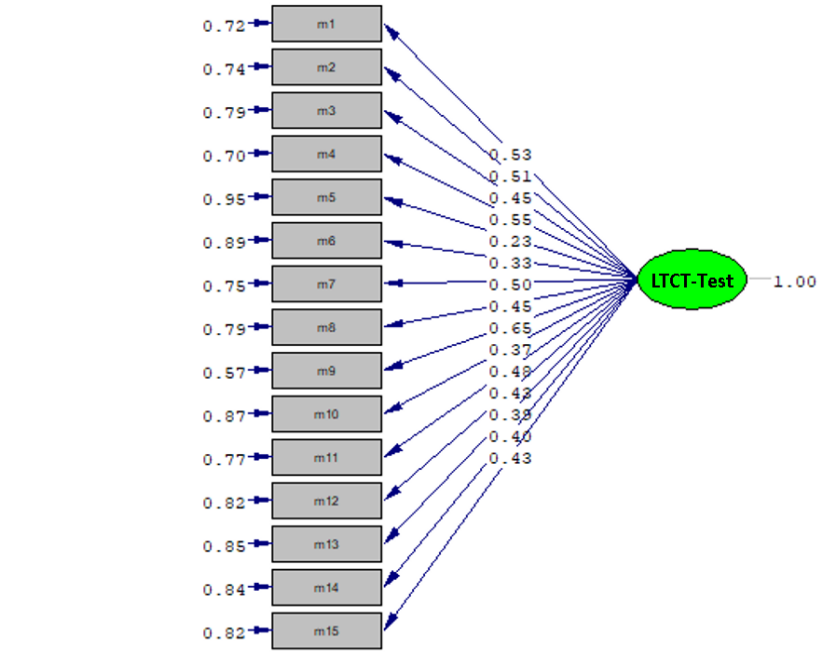


Figure 5. *CFA Diagram for standardized one-factor model solution of LTCT-Test*

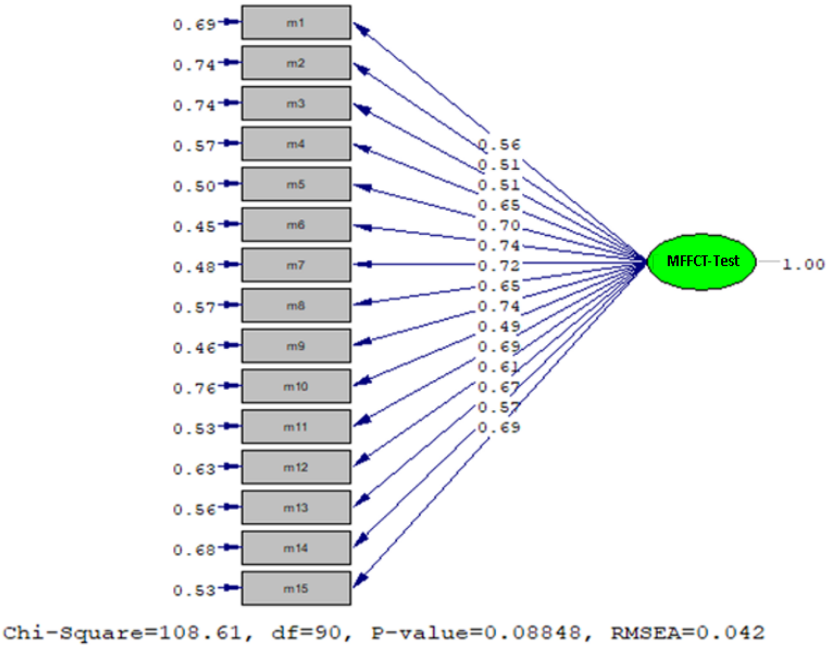


Figure 6. *CFA diagram for standardized one-factor model solution of MFFCT-Test*

The model fit indices are presented in Table 5 according to the results of CFA analysis. The critical value for χ2statistic is considered with degrees of freedom because of its sensibility to sample size. The value of $\chi 2/df < 3$ indicates the perfect fit (Kline, 2015; Tabachnick & Fidell, 2013). The Root Mean Square Error of Approximation value of (RMSE $\leq$ .08) and Standardized Root Mean Square Residual (SRMR $\leq$ .08) are considered as an acceptable fit indicator (Brown, 2015; Hair et al., 2009). According to Kline, approximation of the goodness of fit index values of Comparative Fit Index (CFI $\geq$ .95) and Non-Normed Fit Index (NNFI $\geq$ .95) indicate a good fit. As regards to these criteria, it can be interpreted that both LTCT-Test and MFFCT-Test indicated a very good fit with the one-factor model.

Table 5. The model-fit statistics for LTCT-Test and MFFCT-Test

Test	df	χ2	χ2/df	RMSEA	CFI	NNFI/TLI	SRMR
LTCT-Test	90	121.14	1.346	.049	.96	.95	.066
MFFCT-Test	90	108.61	1.206	.042	.99	.98	.051

3.5. Pre- and Post-test Comparisons of the Instruments

The mean values of the tests were evaluated by considering the total scores of the test and the three elements of the critical thinking skills including interpretation, analysis, and inference scores. The LTCT-Test and MFFCT-Test scores of the students are presented in Figure 7 and Figure 8. As seen in Figure 7, the mean values of the pre-test and post-test of LTCT-Test were 12.00 and 23.22 with the standard deviation of 6.65 and 9.83, respectively. As seen in Figure 8, the mean values of the pre-test and post-test of MFFCT-Test were 13.09 and 23.09 with the standard deviation of 7.98 and 10.02, respectively.

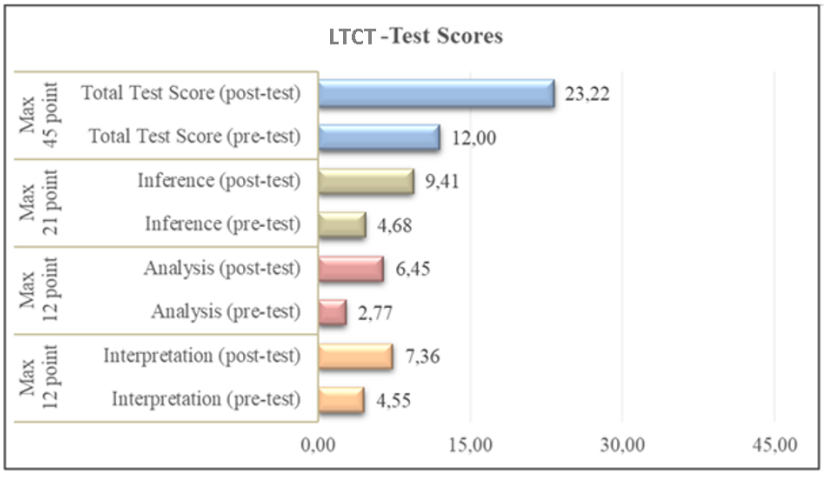


Figure 7. *Mean values for LTCT-Test*

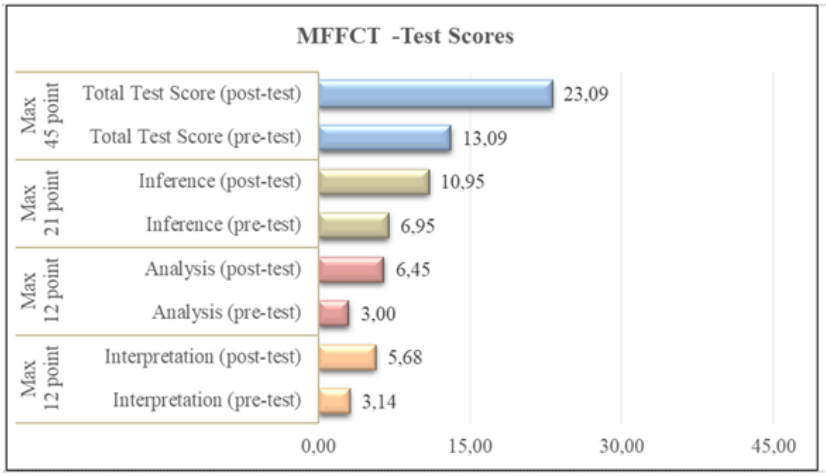


Figure 8. *Mean values for MFFCT- Test*

Then, Figure 7 and Figure 8 indicates that post-test total scores means, and the elements of the critical thinking skills of interpretation, analysis, and inference scores means for both tests are higher than pre-test scores' means. In order to determine whether there were any statistically significant mean differences in the pre- and post-test scores, the Wilcoxon test was calculated. The results for LTCT-Test and MFFCT-Test are presented in Table 6 and Table 7, respectively.

Table 6. *Wilcoxon test results of LTCT-Test*

LTCT- Test Scores	Ranks	*N*	Mean Rank	Sum of Ranks	Z	*P*
Total	Negative Ranks	1	1.00	1.00	-3.982	.000
	Positive Ranks	20	11.50	230.00		
	Ties	1				
	Total	22				
Inference	Negative Ranks	1	5.50	5.50	-3.940	.000
	Positive Ranks	21	11.79	247.50		
	Ties	0				
	Total	22				
Analysis	Negative Ranks	4	4.00	16.00	-3.598	. 000
	Positive Ranks	18	13,17	237.00		
	Ties	0				
	Total	22				
Interpretation	Negative Ranks	3	3.67	11.00	-3.271	.000
	Positive Ranks	15	10.67	160.00		
	Ties	4				
	Total	22				

Regarding the Wilcoxon test results presented in Table 6 there were significant mean differences in total test score of LTCT-Test (Z =-3.982, p = .00 <.05) and sub-skills of inference (Z = -3.940, p = .00 <.05), analysis (Z = -3.598, p = .00 <.05) and interpretation (Z =-3.271, p = .00 <.05) between the pre- and post-test applications. It can be concluded that the designed and implemented living things module based on STEM education approach was an effective way to develop the critical thinking skills of the participant students.

Table 7. *Wilcoxon test results of MFFCT-Test*

MFFCT-Test Scores	Ranks	*N*	Mean Rank	Sum of Ranks	Z	*P*
Total Test Score	Negative Ranks	1	1.00	0	-4.076	.000
	Positive Ranks	21	12.00	252.00		
	Ties	0				
	Total	22				
Inference	Negative Ranks	2	5.00	5.50	-3.861	.000
	Positive Ranks	21	11.79	226.50		
	Ties	0				
	Total	22				
Analysis	Negative Ranks	2	2.00	3.00	-3.818	. 000
	Positive Ranks	19	10.94	207.00		
	Ties	1				
	Total	22				
Interpretation	Negative Ranks	4	2.75	11.00	-3.392	.000
	Positive Ranks	15	11.93	179.00		
	Ties	3				
	Total	22				

Regarding the Wilcoxon test results presented in Table 7 there were significant mean differences in the total test scores of MFFCT-Test (Z = -4.076, p = .00 <.05) and sub-skills of inference (Z = -3.861, p = .00 <.05), analysis (Z = -3.818, p = .00 <.05) and interpretation (Z = -3.392, p = .00 <.05) between the pre- and post-test applications. It can be concluded that the designed and implemented force and friction module based on STEM education approach was an effective way to develop the critical thinking skills of the participant students.

4. DISCUSSION and CONCLUSION

The present study aimed to develop and validate science critical thinking skill instruments to assess the improvement in the subject of living organisms and force and friction through design-based STEM Education. In doing so, LTCT-Test and MFFCT-Test consisting of two-tier 15 multiple-choice items were developed by integrating related science content and three sub-skills of critical thinking as interpretation, analysis and inference with reference to Facione's (1990) Delphi study. In the initial phase of the instruments, 19 two-tiers items for LTCT- Test and 20 two-tier items for MFFCT-Test were written by considering the objectives of the science curriculum (MoNE, 2018) and sub-skills of critical thinking.

For content validity, both test items were evaluated by a group of experts considering the relevance of the item with the content, construct, grade level, and clarity. The modified kappa values were calculated for each item to test interrater reliability and the items with the value less than 0.60 were deleted After the minor revisions on wording in the retaining items, the 15 two-tier item LTCT-Test was applied to 147 students and the 15 two-tier item MFFCT-Test was applied to 116 students at the pilot stage. The item analysis for the first tier of dichotomous items were carried out by using item difficulty (p) and point biserial coefficient (r_{pb}) for both LTCT-Test and MFFCT-Test. The results of item analysis with TAP program pointed out that the values were in an acceptable range. Then, the tests had the average difficulty and discrimination index. In addition, reliability analysis of *KR-20* for the first tier of the tests produced the acceptable value of .759 and .767 for LTCT-Test and MFFCT-Test, respectively. Also, reliability analysis of Cronbach's alpha coefficient for entire tests as polytomous form produced the acceptable value of .789 and .908 for LTCT-Test and MFFCT-Test, respectively.

The second tier of items was evaluated by a polytomous rubric and then total scores were calculated for each item by summing the scores obtained from the first tier and second tier. Therefore, the parallel analysis was carried out based on PCM to determine the number of dimensions and the structure of tests. The results of the parallel analysis suggested extracting only one factor structure for LTCT-Test and MFFCT-Test. Further analysis to confirm the unidimensionality of the CFA was carried out. The results of the CFA confirmed the one factor structure of the tests. In other words, exploratory and confirmatory factor analysis supported the model as a one-dimensional measure for both LTCT-Test and MFFCT-Test with very good fit indices.

In conclusion, when the findings of the content and construct validity, item, and reliability analysis are considered, all items of both tests are valid to measure the critical thinking skills in the related science content as unidimensional. Much of the recent studies examining critical thinking skills have converged on the need for domain-specific teaching and assessment (Asysyifa et al., 2019; Mabruroh & Suhandi, 2017; Mapeala & Siew, 2015; Reynders et al. 2020; Sadhu & Laksono, 2018; Sya'bandari et al., 2017). In this respect, developing and validating instruments to assess the integrated critical thinking skills will contribute to the empirical examination of this construct within the context of school science learning.

The second phase of the research focused on assessing the improvement in students' integrated critical thinking skills in the subject of living organisms and force and friction through design-based STEM education. As an intervention, students participated in STEM modules enriched with critical thinking principles. To do this, the elements of critical thinking skills were reflected in the worksheets and emphasized during the implementation of the STEM modules by the teacher. Both at the beginning and at the end of the modules LTCT-Test and MFFCT-Test were conducted as pre-and post-tests. The collected data were computed with Wilcoxon test. The results were found statistically significant. In other words, participating in the STEM modules enriched with critical thinking principles improved the students' critical thinking skills such as interpretation, analysis, and inference in relation to the science content. This result is consistent

with the previous literature in that increasing the critical thinking skills of students have been found positively related to STEM studies (Baharin et al., 2018; Duran & Şendağ, 2012; Mutakinat et al., 2018; Oonsim & Chanprasert, 2017; Rahmawati et al., 2019; Waddell, 2019).

As a result of this study it can be concluded that the infusing approach is an efficient way to teach critical thinking skills to students through the implementation of the science units. From this point of thought, (Willingham, 2008) suggests that thinking critically should be taught in the context of subject matter and opportunities must be given to students on their own ways to think critically. Further investigations should be required to understand instructional effectiveness and classroom dynamics to contribute designing a more effective educational environment and measure students' critical thinking skills by utilizing both qualitative and quantitative research techniques. Given the significant role of critical thinking skills in nurturing successful individuals in their daily life, teachers ought to be equipped with effective principles and strategies that enable them to sustain student engagement in critical learning activities.

Acknowledgements

This study includes a part of doctoral dissertation entitled Design, Implementation, and Evaluation of the Fifth Grade Science Course Units with an Integrated STEM Education Approach (Doğan, 2020).

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 author(s). For the research, the necessary ethical approval was taken from the ethical committee and the necessary permission was obtained according to board decision, dated 05/02/2019 and 2439445 numbered from the Research Evaluation and Investigation Committee of the Antalya National Education Directorate. Written informed parental consent form was obtained for all students.

ORCID

Ayse Savran Gencer https://orcid.org/0000-0001-6410-152X

Hilmi Dogan https://orcid.org/0000-0001-7933-4115

5. REFERENCES

Abdallah, M. M. S. & Wegerif, R. B. (2014). *Design-based research (DBR) in educational enquiry and technological studies: A version for PhD students targeting the integration of new technologies and literacies into educational contexts.* ERIC: ED546471. Retrieved 3, 2019, from http://files.eric.ed.gov/fulltext/ED546471.pdf

Arıcak, O. T., Avcu, A., Topçu, F., & Tutlu, M. G. (2020). Use of item response theory to validate cyberbullying sensibility scale for university students. *International Journal of Assessment Tools in Education, 7*(1), 18-29. Retrieved October 3, 2020, from https://dx.doi.org/10.21449/ijate.629584

Asysyifa, D.S., Jumadi, Wilujeng, I., & Kuswanto, H. (2019). Analysis of students critical thinking skills using partial credit models (PCM) in physics learning. *International Journal of Educational Research Review, 4*(2), 245-253. https://doi.org/10.24331/ijere.518068

Awang, Z. (2012). *A handbook on structural equation modeling using AMOS* (6th Ed). Universiti Teknologi Mara Press: Malaysia.

Baharin, N., Kamarudin, N., & Manaf, U. K. A. (2018). Integrating STEM education approach in enhancing higher order thinking skills. *International Journal of Academic Research in*

Business and Social Sciences, 8(7), 810–822. Retrieved February 3, 2019, from http://dx.doi.org/10.6007/IJARBSS/v8-i7/4421

Bailin, S. (2002). Critical thinking and science education. *Science & Education*, *11*(4), 361-375. Retrieved February 3, 2019, from http://dx.doi.org/10.1023/A:1016042608621

Beyer, B. K. (1995). *Critical thinking.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Brooks, G. P. & Johanson, G. A. (2003). TAP: Test analysis program. *Applied Psychological Measurement*, *27*(4), 303-304.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Bursal, M. (2017). *SPSS ile temel veri analizleri.* Ankara: Anı Yayıncılık.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı [Data analysis handbook for social sciences]*. Ankara: Pegem Akademi.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London: Sage Publications.

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

Çorlu, M. S., Capraro, R. M. & Capraro, M. M. (2014). Introducing STEM education: Implications for educating our teachers for the age of innovation. *Education and Science, 39*(171), 74-85.

Deng, N., Wells, C., & Hambleton, R. (2008). A confirmatory factor analytic study examining the dimensionality of educational achievement tests. *NERA Conference Proceedings 2008*. 31. Retrieved January 10, 2020, from https://opencommons.uconn.edu/nera_2008/31

Dilekli, Y. (2019). *Etkinliklerle düşünme eğitimi*. Ankara: Anı Yayıncılık.

Doğan, N. (2019). *Eğitimde ölçme ve değerlendirme [Measuremnet and evaluation in education].* Ankara Pegem Akademi.

Doğan, H. (2020). *Design, implementation, and evaluation of the fifth grade science course units with an integrated STEM education approach* (Unpublished doctoral dissertation). Pamukkale University, Denizli, Turkey.

Duran, M., & Şendağ, S. (2012). A preliminary investigation into critical thinking skills of urban high school students: Role of an IT/STEM program. *Creative Education*, *3*(2), 241–250. Retrieved February 3, 2019, from http://dx.doi.org/10.4236/ce.2012.32038

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs: Prentice-Hall.

Ennis, R. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4-10. Retrieved March, 3, 2017 from https://doi.org/10.3102/0013189X018003004

Ergün, A. & Külekci, E. (2019). The effect of problem based STEM education on the perception of 5th grade students of engineering, engineers and technology. *Pedagogical Research, 4*(3), em0037. Retrieved March, 3, 2020 from https://doi.org/10.29333/pr/5842

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations* (ERIC Document Reproduction Service No. ED315423). Retrieved March, 3, 2017, from https://eric.ed.gov/?id=ED315423

Ferrando, P. J. & Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement,* 1–19. https://doi.org/10.1177/0013164417719308

Fraenkel, J.R. & Wallen, N.E. (2008). *How to design and evaluate research in education* (7th ed.). New York: McGraw-Hill.

Gencer, A. S. & Boran, G. H. (2017). *Üst düzey düşünme becerilerinin öğretimi [Teaching higher order thinking skills*]. In S. Dal & M. Köse (Ed), Öğretim ilke ve yöntemleri (pp. 405-445). Ankara: Anı Yayıncılık

Griffard, P. B., & Wandersee, J. H. (2001). The two-tier instrument on photosynthesis: what does it diagnose? *International Journal of Science Education, 23*(10), 1039-1052. Retrieved March, 3, 2017, from https://doi.org/10.1080/09500690110038549

Hair, J. F., Black, W. C, Babin, B.J., & Anderson, R. E. (2009). Multivariate data analysis (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Haladyna, T. M. (1994). *Developing validating multiple choice test items.* Lawrence Erlbaum Associates, Publishers.

Harjo, B., Kartowagiran, B., & Mahmudi, A. (2019). Development of critical thinking skill instruments on mathematical learning high school. *International Journal of Instruction, 12*(4), 149-166. Retrieved January 10, 2020, from https://doi.org/10.29333/iji.2019.12410a

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–64.

Honey, M., Pearson, G., & Schweingruber, H. (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. Washington, DC: National Academies Press.

Kline, B. R. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kuş, M. & Çakıroğlu, E., (2020). Prospective mathematics teachers‟ critical thinking processes about scientific research: Newspaper article example. *Turkish Journal of Education, 9*(1), 22-45. https://doi.org/10.19128/turje.605456

Mabruroh, F. & Suhandi, A. (2017). Construction of critical thinking skills test instrument related the concept on sound wave. *IOP Conference Series: Journal of Physics: Conf. Series* 812 (2017) 012056 https://doi.org/10.1088/1742-6596/812/1/012056

Mapeala, R. & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth graders. *Springer Plus, 4(*741). DOI 10.1186/s40064-015-1535-0

Massachusetts Department of Education. (2006). *Massachusetts science and technology/engineering curriculum framework*. Retrieved January 10, 2019, from http://www.doe.mass.edu/frameworks/scitech/1006.do

Ministry of National Education. (2018). *Elementary and middle school (3, 4, 5, 6, 7, and 8th grades) science curriculum*. Ankara: Board of Education and Training.

Mutakinati, L., Anwari, I., & Yoshisuke, K. (2018). Analysis of students' critical thinking skill of middle school through stem education project-based learning. *Journal Pendidikan IPA Indonesia, 7*(1), 54-65. Retrieved February 3, 2019, from http://journal.unnes.ac.id/index.php/jpii https://doi.org/10.15294/jpii.v7i1.10495

NGSS Lead States. (2013). *Next generation science standards: For states by states.* Washington, DC: The National Academies Press.

National Research Council. (2011). *Assessing 21st Century Skills: Summary of a Workshop.* Washington, DC: The National Academies Press.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill, Inc.

Öner, A. T., Navruz, B., Biçer, A., Peterson, C. A., Capraro, R.M., & Capraro, M.M. (2014). T-STEM academies‟ academic performance examination by education service centers: A Longitudinal Study. *Turkish Journal of Education, 3*(4), 40-51.

Oonsim., W. & Chanprasert, K. (2017). Developing critical thinking skills of grade 11 students by STEM education: Focus on electrostatic in physics. *Rangsit Journal of Educational Studies, 4* (1), 54-59.

Patton, M.Q. (2014). *Qualitative research and evaluation methods: Integrating theory and practice.* Thousand Oaks, CA: Sage Publications.

Polit, D. & Hungler, B., P. (1999). *Nursing research: Principles and methods*. Philadelphia: Lippincott Company.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459-467.

Rahmawati, Y., Ridwan, A., Hadinugrahaningsih, T., & Soeprijanto (2019). Developing critical and creative thinking skills through STEAM integration in chemistry learning. *IOP Conference. Series: Journal of Physics: Conf. Serie*s 1156 (2019) 012033. https://doi.org/10.1088/1742-6596/1156/1/012033

Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *International Journal of STEM Education, 7*(9). Retrieved February 3, 2019, from https://doi.org/10.1186/s40594-020. 00208-5

Sadhu, S. & Laksono, E. W. (2018). Development and validation of an integrated assessment for measuring critical thinking and chemical literacy in chemical Equilibrium. *International Journal of Instruction, 11*(3), 557-572. Retrieved February 3, 2019, from https://doi.org/10.12973/iji.2018.11338a

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. New Jersey: Lawrance Erlbaum Association, Inc.

Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2008). *Thinking-based learning: Promoting quality student achievement in the 21st century*. New York, NY: Teachers College Press.

Sya'bandari, Y., Firman, H., & Rusyat, L. (2017). The development and validation of science virtual test to assess 7th grade students' critical thinking on matter and heat topic. *Journal of Science Learning, 1*(1), 17-27.

Timmerman, M. E. & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220. Retrieved February 3, 2019, from https://doi.org/10.1037/a0023353

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

Waddell, B. (2019). Influence of STEM lessons on critical thinking (Unpublished master's thesis). The Graduate College at the University of Nebraska, Lincoln. Retrieved February 3, 2019, from https://digitalcommons.unl.edu/teachlearnstudent/103

Wang, H. H., Moore, T. J., Roehrig, G. H., & Park, M. S. (2011). STEM integration: Teacher perceptions and practice. *Journal of Pre-College Engineering Education Research, 1*(2), 1-13. Retrieved February 3, 2019, from https://doi.org/10.5703/1288284314636

Whiston, S. C. (2012). *Principles and applications of assessment in counseling* (4th ed.). Belmont, CA: Brooks/Cole, Cengage Learning.

Wicaksana, Y. D., Widoretno, S., & Dwiastuti, S. (2020). The use of critical thinking aspects on module to enhance students' academic achievement. *International Journal of Instruction, 13*(2), 303-314. Retrieved February 3, 2019, from https://doi.org/10.29333/iji.2020.13221a

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction* (8th ed.). Boston: Pearson/A and B.

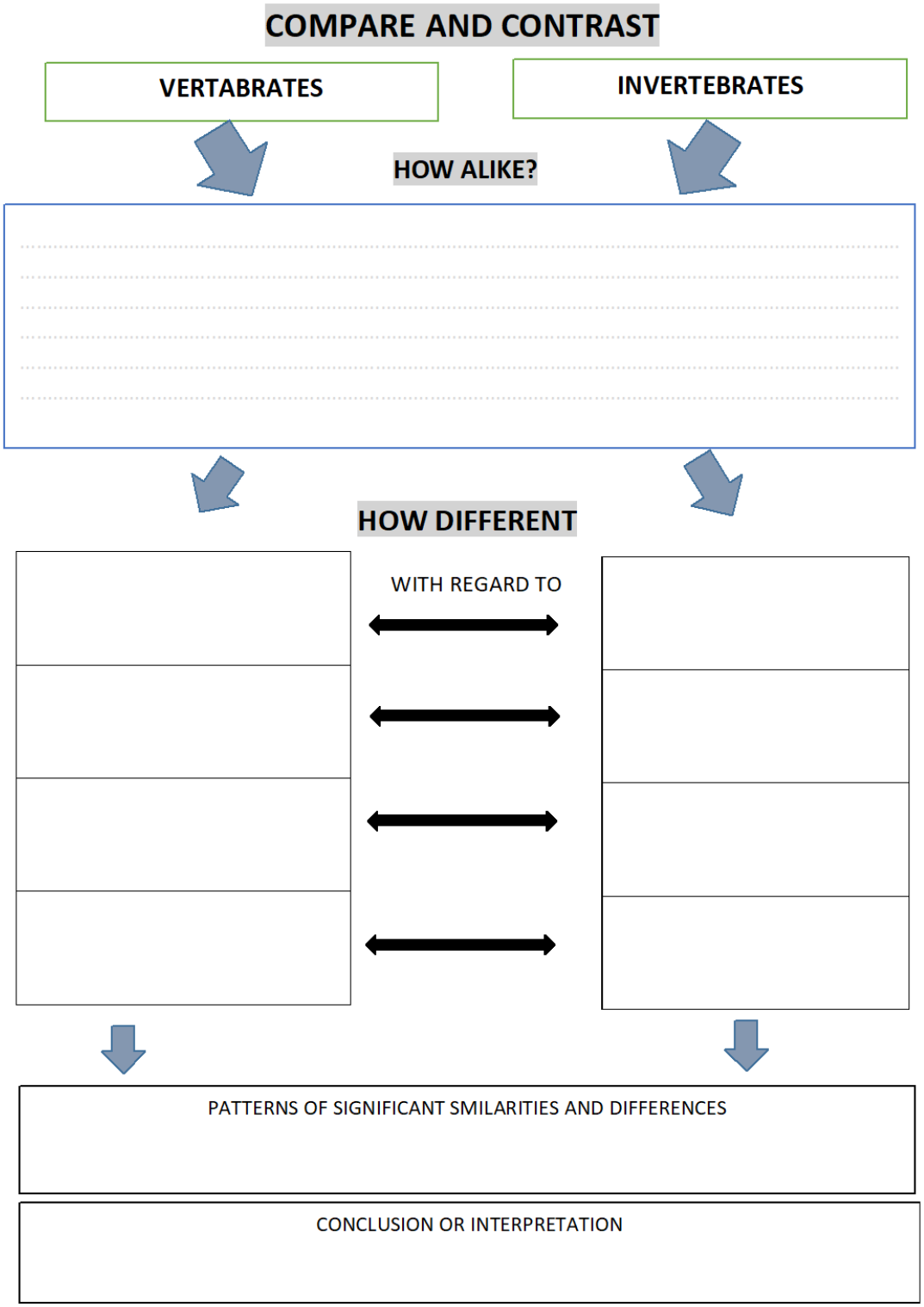
Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, *109*(4), 21-32. Retrieved February 3, 2019, from https://doi.org/10.3200/AEPR.109.4.21-32

Yurdugül, H. & Bayrak, F. (2012). Content validity measures in scale development studies: Comparison of content validity index and kappa statics. *Hacettepe University Journal of Education, Special Issue 2,* 264-271.

Yalaki Y., Doğan, N., İrez, S., Doğan, N., Çakmakçı, G., Kara, B. E. (2019). Measuring nature of science views of middle school students. *International Journal of Assessment Tools in Education, 6(3),* 461-475. Retrieved October 3, 2019, from https://dx.doi.org/10.21449/ijate.561154

APPENDIX

Appendix I. Examples of worksheets for the integrated critical thinking skills.



Adapted from Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2008). *Thinking-based learning: Promoting quality student achievement in the 21st century*. New York, NY: Teachers College Press.

MAKE A DECISION

Research Question: Are there other organisms other than plants and animals?

What did you see under the microscope? Draw the shape below.

Paramecium	Euglena
FEATURES	
Can't make its own food Absorb food its environment It has no real colour Can move in water Can change its shape Can reproduce	Can make its own food by photosynthesis Absorb food its environment It is green Can move in water Has a flagellum Reacts to light It has an eye called stigma Can change its shape Can reproduce Sensitive to temperature
DECISION MAKING	
Evidences that it is an animal	
Evidences that it is a plant	
Evidences that it is both a plant and an animal	
Do you think these organisms are animals or plants? Or should it be in another group? Explain why you think so by writing your decision.	

Adapted from Osborne, J., Erduran, S., & Simon, S. (2004). *Ideas, evidence and argument in science. Video, in-service training manual and resource pack*. London: King's College London.

DETERMINING PARTS -WHOLE RELATIONSHIP

THE WHOLE OBJECT

PARTS OF A DYNAMOMETER (What are the parts made of?)

Spring	Hook	Tube	Graduated scale

HOW DYNAMOMETER WILL WORK IF THE PARTS OF DYNAMOMETER WOULD BE BROKEN OR MISSING?

Explain how these pieces work together:

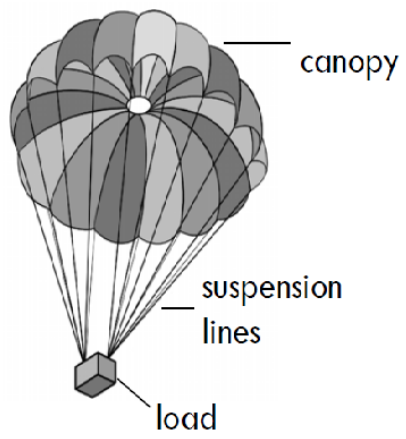
Adapted from Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2008). *Thinking-based learning: Promoting quality student achievement in the 21st century*. New York, NY: Teachers College Press.

PROBLEM SOLVING

Problem: Aerospace engineers want to land a robot on a newly discovered planet with a parachute. However, the atmosphere of this planet is thinner than the atmosphere of the Earth. It is known that the planet's gravitational force is greater than the Earth.

Design a parachute that can safely deliver the robot to the surface of this new planet.

Possible Solutions
How can I solve the problem?



A parachute that works well on the Earth

To solve the problem, which modifications should you make on a parachute that works well on the Earth? Please explain.

Write more than one solution suggestion.
Solution 1 ……

Which solution should I choose?

What may happen if you follow this solution ?	Pros and Cons	How important is this result? Why?

New solution proposal
How can I improve my solution to solve this problem?

Adapted from Engineering is Elementary (n.d). *Designing Parachutes*. Museum of Science, Boston. Retrieved from http://d7.eie.org/sites/default/files/resource/file/pa_student_assessments.pdf