Development of Science Achievement Test Including the Units of "States of Matter and Heat" and "Electricity in Our Life"

Serkan Timur [i]
Çanakkale Onsekiz Mart University

Eylem Yalçınkaya Önder [ii]
Çanakkale Onsekiz Mart University

Betül Timur [iii]
Çanakkale Onsekiz Mart University

Meltem Ekici [iv]
Çanakkale Onsekiz Mart University

Abstract

In this study, a valid and reliable multiple-choice achievement test has been developed in order to measure the academic achievement of 8th grade students about the units of "States of Matter and Heat" and "Electricity in Our Life". The pilot study of the achievement test was carried out with 30 questions which were prepared with at least one question from each gain related to unit gains. This study was conducted with 287 students having learned these subjects from two different schools in the central district of Siirt in 2017-2018 academic year. The validity and reliability studies of the achievement test items were performed. For the validity of the test, a table of specifications was prepared, and the test was examined by two faculty members and three science teachers. Item analysis and reliability analyzes of the test were performed with TAP 14.7. As a result of the analyzes, 3 items with low item discrimination index were excluded from the test. Mean discrimination index of the final test including 27 questions was found to be 0.520 and mean item difficulty index was 0.598. As a result of reliability analysis; KR_{20} value was calculated as 0,840 and Spearman-Brown value was 0,730. As a result of the study, a valid and reliable multiple-choice achievement test was introduced to the science education to measure the academic success of the students in the 8th grades about the units of "States of Matter and Heat "and "Electricity in Our Life".

Keywords: Science, Education, Achievement Test, Electricity, States of Matter, Heat, 8th grade

DOI: 10.29329/ijpe.2020.228.1

---

[i] **Serkan Timur,** Assoc. Prof. Dr., Fen Bilgisi Eğitimi Abd, Çanakkale Onsekiz Mart Üniversitesi, ORCID: 0000-0002-4949-2275

[ii] **Eylem Yalçınkaya Önder,** Assist. Prof., Fen Bilgisi Eğitimi Abd, Çanakkale Onsekiz Mart Üniversitesi, ORCID: 0000-0003-1306-9931

Correspondence: eylemyk@gmail.com

[iii] **Betül Timur,** Assoc. Prof. Dr., Fen Bilgisi Eğitimi Abd, Çanakkale Onsekiz Mart Üniversitesi, ORCID: 0000-0002-2793-8387

[iv] **Meltem Ekici,** Lecturer, Fen Bilgisi Eğitimi Abd, Çanakkale Onsekiz Mart Üniversitesi

INTRODUCTION

Education is defined as '*The process of bringing about the desired behavior change in one's own behavior through his own experience*' (Ertürk, 1975). It is important to know to what extent the individual is successful in education, which is the process of emergence of changes in the individual, and to what extent changes in behavior are (Erdoğdu & Kurt, 2012). For this purpose, it is important to determine whether the behavior change occurs in the student or not and whether the achievement of the desired goals is achieved (Ertürk,1975; Baykul, 2000; Gömleksiz & Erkan, 2016).

Assessment and evaluation has the characteristics such as the extent to which the teaching has reached the objectives, the suitability of the subject to the student's abilities, guidance to the next processes, providing the necessary feedback to support the teaching, feedback and informing about the progress of the teaching process (Gömleksiz & Erkan, 2016; Küçük & Geçit, 2012; Birgin, 2010; Gümüş, 1977).

Teachers have an important role in measurement and evaluation, education and training activities due to reasons such as how much of the achievements to be achieved in schools in line with teaching purposes and the determination of measurement errors and deficiencies in the process (Küçük & Geçit, 2012). Behaviors that require the most measurements in the education process are those belonging to the cognitive field. Cognitive behaviors and learning are usually related to mental abilities and activities. They consist of behaviors such as defining, interpreting, assimilating, applying and generalizing information (Gömleksiz & Erkan, 2016).

The various questions used in assessment and measurement are one of the most important tools to determine the extent to which the individual has achieved the goals and objectives. Since the traditional approach to education has been replaced by the constructivist approach, the evaluation of the process has come to the forefront. However, it is accepted that the exam and question types brought by the traditional method can be used for a long time and will continue to be widely acclaimed (Küçük & Geçit, 2012).

Measurement and evaluation are of great importance in all activities to be carried out for education and training. Therefore, measurement tools should be sensitive and reliable according to the quality to be measured (Gömleksiz & Erkan, 2016). In science education, teachers need to use valid and reliable measurement and assessment tools to determine the degree to which an individual achieves goals and achievements correctly (Gönen, Kocakaya & Kocakaya, 2011; Saraç, 2018).

Multiple choice tests are one of the most widely used measurement tools in measuring the simple and complex concepts and providing the opportunity to cover all the gains (Saraç, 2018). They are preferred measurement tools for detecting students' misconceptions, applicability in the class and determining achievement levels on a particular subject (Küçükahmet, 2002; Kan, 2014). However, multiple choice tests are limited in determining students' achievements or abilities in areas of critical thinking skills and creativity (Küçükahmet, 2002). The possibility of marking the correct answer by coincidence without actually knowing the answer is one of the disadvantages of multiple-choice tests (Turgut & Baykul, 2015; Mintzes, Wandersee & Novak, 2001).

There are many studies on multiple choice achievement tests in science education. Some of these studies; concepts in solutions (Çalık & Ayas, 2003), force and motion (Akbulut & Çepni, 2013), simple electrical circuits (Şen & Eryılmaz, 2011), change of matter (Saraç, 2018), heat and temperature (Ayvacı, Hakan, Şevki & Durmuş, 2016), solutions (Demir, Kızılay & Bektaş, 2016), example of cell division and inheritance (Kızılkapan & Bektaş, 2018), structure and properties of matter (Kızkapan & Bektaş, 2018).

The aim of this study is to develop a valid and reliable achievement test covering the gains of 8th grade science course 'States of Matter and Heat' and 'Electricity in our Lives' units. For this purpose, a pilot study of 30 multiple-choice questions selected from various sources was carried out

and 3 items were removed from the test and validity and reliability studies were conducted with 27 items. As a result of the analyzes and the table of specifications, the test was found to be valid and reliable. The test developed in this context is thought to contribute to the related field. At the same time, the achievement test developed by teachers can be used by teachers to measure and evaluate student achievement, to identify missing learning and misconceptions.

METHOD

Survey model as a quantitative research model was used in the study and some steps were followed in the study process. Related literature was reviewed within the scope of 8th grade science course. As a result of the researches, the gains and objectives of the subject matters were determined. Then, following steps are followed to determine the characteristics of the test: the types of items and the writing of the items, item control, pilot study, the analysis of the pilot study results, and the final version of the test was formed (Gömleksiz & Erkan, 2016; Küçük & Geçit, 2012). After these steps, some analysis has been conducted to test the validity and reliability of the achievement test. The reliability of the test was performed in order to measure the reliability of the test, whether it is free of errors or not (Küçükahmet, 2002). After all, a 27-item achievement test was prepared including the gains of the two units.

In the preparation of the items of the achievement test; the Ministry of National Education (MNE) exam questions, MORPA KAMPÜS subject survey tests, OKULİSTİK and "West Academy Science 8th grade test book" were used.

Sample of the Study

A total of 287 students from two secondary schools in the province of Siirt, which had already learned the relevant subjects were participated to this study. These schools were selected from the identified accessible population by the convenience sampling technique. The pilot study was conducted in the spring semester of 2017-2018 academic year.

Data Collection Tools

Achievement test developed by one of the researchers of the study in order to measure the achievement of students about the related subjects. The questions were prepared by taking into consideration the relevant subjects and gains of the Ministry of National Education's (MNE) curriculum in 2017-2018 academic year. The original test consisted of 30 questions. As a result of the pilot study, 3 questions were removed, and the final version included 27 multiple choice questions with four alternatives.

The content validity of the test was examined by the field experts. The test items were based on the MNE curriculum and sources of the test items were presented in the Table 1. Each unites of 'States of Matter and Heat' and 'Electricity in Our Life' were instructed to the students by the teacher for 4 weeks in 2017-2018 academic year. The highest score a student can get from this test was 27.

Table 1. Sources of questions

Question Number	Quoted Sources	Similar Earnings
1	West Academy Publications, 2016-2017	2
2	Morpa Kampüs, 2017-2018	1
3	MNE TEOG Exam, 2016-2017	4
4	MNE TEOG Exam, 2016-2017	3
5 (removed)	Morpa Kampüs, 2017-2018	13
6	Okulistik, 2017-2018 Syf 58	7
7 (removed)	Morpa Kampüs, 2017-2018	6
8	MNE TEOG Exam, 2016-2017	9
9	Morpa Kampüs, 2017-2018	8

10	Morpa Kampüs, 2017-2018	12, 30
11	West Academy Publications, 2016-2017, Pg.124	14, 15
12	Morpa Kampüs, 2017-2018	10, 30
13	West Academy Publications, 2016-2017, Pg.129	5
14	West Academy Publications, 2016-2017, Pg.124	11, 15
15	MNE TEOG Make-up Exam, 2016-2017	11, 14
16	West Academy Publications, 2016-2017, Pg.148	21, 22
17	West Academy Publications, 2016-2017, Pg.145	24
18	MNE TEOG Exam, 2016-2017	23
19	MNE TEOG Make-up Exam, 2016-2017	20
20	MNE TEOG Make-up Exam, 2016-2017	19
21	Morpa Kampüs, 2017-2018	16, 22
22	Morpa Kampüs, 2017-2018	16, 21
23	West Academy Publications, 2016-2017, Pg.141	18
24	West Academy Publications, 2016-2017, Pg.145	17
25	Morpa Kampüs, 2017-2018	26, 29
26 (removed)	Morpa Kampüs, 2017-2018	25, 29
27	West Academy Publications, 2016-2017, Pg.143	28
28	West Academy Publications, 2016-2017, Pg.143	27
29	Okulistik, 2017-2018	25, 26
30	Okulistik, 2017-2018	10, 12

The characteristics measured in achievement tests used in education are generally learning products. Behaviors are the products that result from learnings in schools. Behaviors are also classified among themselves. The most commonly used classifications are those of Bloom and his collaborators (Gömleksiz & Erkan, 2016). The identified behaviors must be associated with the gains to be measured with the objectives. To ensure this, the specification table is created (Demir, Kızılay & Bektaş, 2016). Bloom's taxonomy consists of six cognitive domains consisted of six major categories: knowledge, comprehension, application, analysis, synthesis, and evaluation. The first three of these categories are basic skills, others are considered as high-level thinking skills (Birgin, 2016).

Analysis of Data

TAP (Test Analysis Program) was used for the validity and reliability study of the developed achievement test. Students' answers to the test questions were coded as “A, B, C, D”, and blank answers were deemed as incorrect. Scoring of each correct answer is 1 point. Therefore, the lowest score that can be obtained from the achievement test is “0” and the highest score is “27. Validity is the correspondence between the actual result of the value and purpose the test wants to measure (Gömleksiz & Erkan, 2016). The validity study is whether the measuring instrument is suitable for the intended purpose and whether or not other factors other than those intended are involved in the application (Turgut, 1995; Gömleksiz & Erkan, 2016). For the content validity of the test, first of all a table of specifications was prepared showing the compatibility of the questions with the gains of 2017-2018 MEB Science curriculum. At least two questions were prepared for each gain and three questions have been prepared for some gains. The gains were examined according to the cognitive steps of Bloom's taxonomy. For the validity study, analyzes were performed with TAP, and the discrimination and item difficulty indexes of the items were calculated. The reliability analyzes of the test were found by calculating the KR20 coefficient in the TAP.

FINDINGS AND DISCUSSION

In this section, validity and reliability studies related to achievement test are presented. The content validity of a test must be sufficient to measure all the gains of the targeted subject (Gömleksiz & Erkan, 2016). One of the methods that can be used to ensure content validity is to prepare a table of specifications (Büyüköztürk et al., 2012). In Table 2, the items in which the test occurs respectively in each column, the suitability of these items with the gain and the cognitive step according to revised Bloom's taxonomy were given by the cross.

Table 2. Classification according to Revised Bloom's Taxonomy

Question	Gain	Cognitive Areas					
		Remembering	Understanding	Applying	Analyzing	Evaluating	Creating
1	8.6.1-Define specific heat. He concludes that different substances may have different specific heats as a result of the experiments.	X					
2	8.6.1-Define specific heat. He concludes that different substances may have different specific heats as a result of the experiments.	X					
3	8.6.2.1-Understands the relationship between heat & specific heat, mass & temperature.			X			
4	8.6.2.1-Understands the relationship between heat & specific heat, mass & temperature.				X		
5(removed)	8.6.3.1-He concludes that there is heat exchange during the change of state.		X				
6	8.6.3.4-Associates heat exchanges with change of state in daily life.					X	
7(removed)	8.6.3.4-Associates heat exchanges with change of state in daily life.		X				
8	8.6.3.3-Draws and interprets the change of state graph of the matters.		X				
9	8.6.3.3-Draws and interprets the change of state graph of the matters.		X				
10	8.6.3.2-Calculates and interprets the heat exchange during change of state of substances.			X			
11	8.6.2.1-Solves the problem of heat exchange.				X		
12	8.6.3.2-Calculates and interprets the heat exchange during change of state of substances.			X			
13	8.6.3.1-He concludes that there is heat exchange during the change of state.		X				
14	8.6.2.1-Solves the problem of heat exchange.			X			
15	8.6.2.1-Solves the problem of heat exchange.		X				
16	8.7.1.1-Illustrate and explain electrification by observing the applications in technology and some natural phenomena.	X					
17	8.7.2.2-Knows the intended purpose of the electroscope and shows the working principle.	X					
18	8.7.2.1-Classify the objects in terms of their electrical charges.		X				
19	8.7.1.3-Makes experiments related to the types of electrification and observes the results.				X		
20	8.7.1.3-Makes experiments related to the types of electrification and observes the results.			X			
21	8.7.1.1-Illustrate and explain electrification by observing the applications in technology and some natural phenomena.	X					
22	8.7.1.1-Illustrate and explain electrification by observing the applications in technology and some natural phenomena.		X				
23	8.7.2.1-Classify the objects in terms of their electrical charges.		X				
24	8.7.2.2-Knows the intended purpose of the electroscope and shows the working principle.	X					
25	8.7.2.3-Discover what grounding is and apply it to everyday life and technologies.		X				
26(removed)	8.7.2.3-Discover what grounding is and discuss its importance in terms of safety of life and property by considering its applications in daily life and technology.			X			
27	8.7.1.2-By classifying electrical charges, they discover the effects of the same and different types of electrical charges on each other.			X			
28	8.7.1.2-By classifying electrical charges, they discover the effects of the same and different types of electrical charges on each other.			X			
29	8.7.2.3-Discover what grounding is and discuss its importance in terms of safety of life and property by considering its applications in daily life and technology.			X			
30	8.6.3.2-Calculates and interprets the heat exchange during change of state of substances.	X					

Each item of the achievement test was examined by two faculty members and three science teachers who are experts in their field in terms of content validity. Based on the results of the analysis, the 5th, 7th and 26th questions with low discrimination index were excluded from the test. The statistical information of the results of the analysis after the questions extracted was presented in Table 3.

Table 3. Achievement test statistical data

Achievement Test	
Number of questions	27
Number of people participated	287
Mean	16.153
Standard deviation	5.641
Skewness	-0.143
Kurtosis	-1.109
Mean item difficulty	0.598
Mean discrimination index	0.520
KR20	0.840
KR21	0.827
Spearman-Brown	0.730
Split-half (1st/ 2nd) reliability	0.575
Mean biserial correlation	0.571

Table 4. Achievement test item analysis

Item number	Number of people	Item difficulty index	Item discrimination index	Upper group correct answer	Subgroup correct answer	Point Biserial correlation	Adj. Ptbis
1	215	0.75	0.32	76	48	0.33	0.26
2	129	0.45	0.33	54	26	0.29	0.20
3	227	0.79	0.51	80	37	0.50	0.45
4	224	0.78	0.41	81	46	0.40	0.34
5	186	0.65	0.40	64	30	0.34	0.26
6	189	0.66	0.50	75	33	0.45	0.38
7	128	0.45	0.65	67	13	0.49	0.42
8	209	0.73	0.48	77	36	0.48	0.42
9	162	0.56	0.70	72	14	0.57	0.51
10	197	0.69	0.37	67	35	0.34	0.27
11	146	0.51	0.41	62	27	0.35	0.27
12	103	0.36	0.36	49	19	0.29	0.21
13	183	0.64	0.33	62	34	0.29	0.21
14	199	0.69	0.56	80	33	0.44	0.37
15	183	0.64	0.81	81	14	0.67	0.62
16	212	0.74	0.65	80	26	0.57	0.51
17	144	0.50	0.65	71	17	0.51	0.44
18	165	0.57	0.65	69	15	0.45	0.38
19	155	0.54	0.75	79	17	0.57	0.51
20	188	0.66	0.71	80	21	0.61	0.55
21	188	0.66	0.62	79	27	0.56	0.50
22	112	0.39	0.49	56	15	0.39	0.32
23	188	0.66	0.58	77	28	0.48	0.41
24	186	0.65	0.54	68	23	0.43	0.36
25	177	0.62	0.45	71	33	0.38	0.30
26	116	0.40	0.46	54	16	0.40	0.33
27	125	0.44	0.36	51	21	0.29	0.20

Item analysis can be performed by following the main steps. First of all, the test was conducted with a group of students by providing sufficient time. After the implementation, the test items of each student were checked, and their total scores were calculated. When the students' scores were ranked from highest to lowest in the calculation process, the first 27% of the test was named as the upper group and the lowest 27% was called the subgroup. The group between these two groups was not included in the item analysis. Item difficulty index and item discrimination index power were calculated for each item of the test (Gönen, Kocakaya and Kocakaya, 2011). Table 4 provides detailed item analysis of all items.

Item difficulty index values are generally between 0.00 and 1.00. If this value is between 1.00 and 0.70, it means the item is very easy, between 0.69 and 0.50 the item is easy, between 0.49 and 0.30 the item is of medium difficulty, and 0.29 and below indicates that the problem is difficult (Küçük & Geçit, 2012). Thus, it can be understood that the number of people answering difficult questions correctly is low and the number of people answering easy questions correctly is high (Gömleksiz & Erkan, 2016; Tekin, 1977). Since the item difficulty index is important for the reliability of the test, the test items need to be of medium difficulty to ensure the high reliability of the test (Gelbal 2004; Çepni et al., 2008, Gömleksiz & Erkan, 2016). When Table 4 was examined, it was seen that most difficulty indices of the items in the test were close to 0.50. Moreover, the mean item difficulty index of all items in the test was found to be 0.598. The fact that the difficulty index of the items was close to 0.50 increased the reliability of the test.

The discriminative power index of item is the degree of sort out of the person who knows from the person who does not or the degree of sort out of successful students from unsuccessful ones. As the discriminative power of the item increases, so does its reliability. Therefore, the test items should have high discriminative power index. Item discrimination power can take any value between 0.00 and 1.00. When the discriminating power approaches to zero, it means that the degree of sorting out of successful students from unsuccessful ones is low. When it approaches to 1 indicates that the degree of separation is high. If the discriminative index of the item is negative, it means unsuccessful students or students with low scores on the test answered that test item more correctly than students with high scores. These items must be removed from the test (Gömleksiz & Erkan, 2016; Küçük & Geçit, 2012; Atılgan, 2006; Tekindal 2009). According to Tekin (2003), an item discrimination index of 0.40 or more indicated it is a very good item, between 0.39 and 0.30 means a guide a good item and between 0.29 and 0.20 needs to be improved and corrected, 0.19 and smaller is very weak and needs to be removed from the test. In the current study, the 5^{th}, 7^{th} and 26^{th} questions having 0.26, 0.23, 0.07 item discrimination indexes respectively were excluded from the test since they had a negative effect on reliability (See Table 4). After the removal of these 3 test items, the analyzes were repeated and it was found that the item discrimination power index of all items was greater than 0.30. The fact that the mean item discrimination index of the 27 items was found as 0.520 which has a positive effect on the reliability of the test. Therefore, it can be said that all the items in the final state of the test were functional. In addition, item-total correlation shows the relationship between the total score of all test participants and the score of a single question. The positive and high value of the item-total correlation increases the internal consistency of the test (Büyüköztürk, 2010). The item correlation coefficient should be positive and take a value of 0.20. If this value is negative or low, it indicates that the item/question measures a different property than the other items/questions. (Şener & Taş, 2017). When the Biserial correlation coefficient of all items in the test was considered, it was seen that it is greater than 0.20. The mean biserial correlation coefficient of the test was 0.571. The high correlation between the items indicates that the items are homogeneously distributed in the achievement test (Tavşancıl, 2006).

Another feature that should be in the achievement tests is reliability. For this purpose, internal consistency coefficient and Spearman-Brown correlation coefficient were calculated with the help of TAP 14 program. The internal consistency coefficient, which provides information about the compatibility between the substances after the test, is generally appropriate to calculate for one-dimensional tests (Ebel 1972). If the test is consistent in itself, it becomes possible for the test to produce reliable results (Gömleksiz ve Erkan, 2016). KR20 and KR21 can be used in tests with item analysis. KR20 is used to determine the internal consistency of the test results. KR21 can be used if the test items have similar difficulties in item analysis (Büyüköztürk, 2004). The reliability coefficient can be between 0.00 and 1.00. This value cannot have a negative value. Tests with a reliability coefficient of 0.70 and above are considered reliable (Fraenkel & Wallen, 2009). As a result of the analyzes, the KR20 value of the 30-question test was 0.831, and the value of KR20 increased to 0.840 after the exclusion of 3 items. The reliability of the test was calculated by Sperman-Brown test by two equivalent halving methods and this value was found as 0.730. Based on the K20 value and Sperman-Brown test results, it can be said that this achievement test is reliable.

CONCLUSION

Multiple choice tests are one of the most common tools used for measurement and evaluation in educational systems. Although they have limitations in measuring creativity, they are generally used to measure achievement (Küçükahmet, 2002; Demir, Kızılay & Bektaş, 2016).

Many achievement tests have been developed in the field of science. However, most of these tests include the gains of a single unit or subject. However, especially in educational institutions, the achievements of a single unit are not always measured and, in some cases, exams involving two and more units need to be prepared. The fact that the developed test includes all the gains of the two units will also contribute to the field.

The validity studies of the prepared test were conducted, the table of specifications of the test was prepared and the questions were prepared with the help of expert opinions. In this context, the test was found to be appropriate to the level of the students and the purpose of the test and the achievements of the subjects. As a result of item analysis, 3 questions were excluded from the test and achievement tests' item discrimination index was found to be 0.520 and item difficulty power index was found to be 0.598. The fact that the item discrimination index was greater than 0.40, and the item difficulty index was of medium difficulty had a positive effect on ensuring the validity and reliability of the test (Tekin, 2003; Gelbal, 2004).

As a result of the reliability analyzes of the test, KR_{20} value of the test was found to be 0,840 and Sperman-Brown value was found to be 0,730. These values demonstrated that this developed achievement test is reliable.

Students' achievement and missing learnings about the related subjects can be determined with the help of the developed achievement test. Achievement test can be used to measure the achievement level of the related science course and the other studies to be conducted in the field of science.

Suggestions

Based on the findings of the current study, the following suggestions can be made:

- Since the achievement test developed is a multiple-choice test, it may be inadequate in some comprehensive studies. Using different measurement and evaluation techniques in such studies would increase the validity and reliability of the test.
- The achievement test was based on the 2017-2018 curriculum. Therefore, it may not meet the other years of curricula. In this case, the use of additional measurement tests would increase the validity and reliability of the tests.
- This achievement test can be used to determine students' pre-learning and achievement of objectives of the students on the related subject.
- This test may be incomplete in measuring high-level cognitive domains. The scope of the study can be expanded by carrying out studies on this subject.

REFERENCES

Akbulut, H. İ., & Çepni, S. (2013). Bir Üniteye Yönelik Başarı Testi Nasıl Geliştirilir?: İlköğretim 7. Sınıf Kuvvet ve Hareket Ünitesine Yönelik Bir Çalışma. *Amasya Üniversitesi Eğitim Fakültesi Dergisi*, *2*(1), 18-44.

Atılgan, H. (2006). Ölçme ve Değerlendirme. Öğretmen Adayları İçin Tamamı Konu Anlatımlı Eğitim Bilimleri KPSS. (Sönmez, V. ed.). *Ankara: Çağdaş Öğretmen Yayınları.*

Ayvacı, H. Ş., & Durmuş, A. (2016). Bir Başarı Testi Geliştirme Çalışması: Isı ve Sıcaklık Başarı Testi Geçerlik ve Güvenirlik Araştırması. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi*, *35*(1), 87-103.

Baykul, Y., (2000). Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması. *Ankara, ÖSYM yayınları, 141*.

Birgin, O. (2016). Bloom Taksonomisi. (Bingölbali, E., Arslan, S., & Zembat, İ. Ö. ed.), Matematik Eğitiminde Teoriler. *Ankara: Pegem Akademi*

Birgin, O. (2010). 4-5. Sınıf Matematik Öğretim Programında Öngörülen Ölçme ve Değerlendirme Yaklaşımlarının Öğretmenler Tarafından Uygulanabilirliği. *Unpublished doctorate dissertation, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü, Trabzon*.

Büyüköztürk, S. (2004). Sosyal Bilimler için Veri Analizi El Kitabı (4th ed.). *Ankara: Pegem Yayıncılık.*

Büyüköztürk, Ş. (2010). Sosyal Bilimler için Veri Analizi El Kitabı (12th ed.). *Ankara: Pegem Akademi.*

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2012). Bilimsel Araştırma Yöntemleri (13th ed.). *Ankara: Pegem Akademi*

Çalık, M., & Alipaşa, A. Y. A. S. (2003). Çözeltilerde Kavram Başarı Testi Hazırlama ve Uygulama. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *14*(14), 1-17.

Çepni, S., Bayrakçeken, S., Yılmaz, A., Yücel, C., Semerci, Ç., Köse, E., Sezgin, F., Demircioğlu, G., & Gündoğdu, K. (2008). Ölçme ve Değerlendirme. *Ankara: Pegem Akademi.*

Demir, N., Kızılay, E., & Bektaş, O. (2016). 7. Sınıf Çözeltiler Konusunda Başarı Testi Geliştirme: Geçerlik ve Güvenirlik Çalışması. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, *10*(1), 209-237.

Ebel, R. (1972). Essentials of Educational Measurement. *New Jersey: Prentice-Hall, Inc. Englewood Cliffs.*

Erdoğdu, M. Y., & Kurt, F. (2012). Öğretmenlerin Ölçme ve Değerlendirme Yeterlik Algılarının Bazı, Değişkenler Açısından İncelenmesi. *Electronic Journal of Education Sciences*, *1*(2), 23-36.

Ertürk, S. (1975). Eğitimde program geliştirme. (2nd ed.). *Ankara: Yelkentepe Yayınları.*

Fraenkle, J. R., & Wallen, N. E. (2009). How to Design And Evaluate Research in Education (7th ed.). *New York:McGrow-Hill.*

Gelbal,S. (2004). Öğretmen Adayları İçin Konu Anlatımlı KPSS. (Demirel, Ö., 20th ed.). *Ankara: Pegem A Yayınları*.

Gömleksiz, M., & Erkan, S. (2016). Eğitimde Ölçme ve Değerlendirme (4th ed.). *Ankara: Nobel Yayın Dağıtım.*

Gönen, S., Kocakaya, S., & Kocakaya, F. (2011). Dinamik Konusunda Geçerliliği ve Güvenilirliği Sağlanmış Bir Başarı Testi Geliştirme Çalışması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *8*(1), 40-57.

Gümüş, B. (1977). Egitimde Ölçme ve Değerlendirme (Measurement and evaluation in education). *Ankara: Kalite Matbaasi.*

Kan, A. (2014). Ölçme Aracı Geliştirme, Eğitimde Ölçme ve Değerlendirme (Satılmış, T. ed.). *Ankara, PegemA.*

Kızkapan, O., & Bektaş, O. (2018). Fen Eğitiminde Başarı Testi Geliştirilmesi: Hücre Bölünmesi ve Kalıtım Örneği. *Maarif Mektepleri Uluslararası Eğitim Bilimleri Dergisi, 2*(1), 1-18.

Küçük, M., & Geçit, Y. (2012). Eğitimde Ölçme ve Değerlendirme (1th ed.). A*nkara: Nobel Yayın Dağıtım.*

Küçükahmet, L. (2002). Öğretimde Planlama ve Değerlendirme (13th ed.). A*nkara: Nobel Yayın Dağıtım.*

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2001). Assessing Understanding in Biology. *Journal of biological education*, *35*(3), 118-124.

Saraç, H. (2018). Fen Bilimleri Dersi 'Maddenin Değişimi' ünitesi ile İlgili Başarı Testi Geliştirme: Geçerlik ve Güvenirlik Çalışması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *18*(1), 416-445.

Şen, H. C., & Eryılmaz, A. (2011). Bir Başari Testi Geliştirme Çalişmasi: Basit Elektrik Devreleri Başari Testi Geçerlik ve Güvenirlik Araştirması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *8*(1), 1-39.

Şener, N., & Taş, E. (2017). Developing Achievement Test: A Research for Assessment of 5th Grade Biology Subject. *Journal of Education and Learning*, *6*(2), 254-271.

Tavşancıl, E. (2006). Tutumların Ölçülmesi ve SPSS ile Veri Analizi (3th ed.). *Ankara: Nobel Publishing.*

Tekin, H. (1977). Eğitimde Ölçme ve Değerlendirme. *Ankara: Mars Matbaası.*

Tekin, H. (2003). Eğitimde Ölçme ve Değerlendirme. (16th ed.). *Ankara: Yargı Yayınevi*

Tekindal, S. (2009). Okullarda Ölçme ve Değerlendirme Yöntemleri. *Ankara: Nobel Yayın Dağıtım.*

Turgut M. F. (1995). Eğitimde Ölçme ve Değerlendirme Metodları. *Ankara: Yargıcı Matbaası.*

Turgut, M. F., & Baykul, Y. (2015). Eğitimde Ölçme ve Değerlendirme. *Ankara*: *Pegem A yayıncılık.*