Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction
12(Special Issue) (2020) 115–128

IJCI
International Journal of Curriculum and Instruction

Developing an adaptation test for students to daily life about the unit "Systems and Health in Our Body"

Miyase TUTAR [a] *, **Orhan KARAMUSTAFAOĞLU** [a]

[a] *Amasya University, Faculty of Education, Amasya and 05100, Turkey*

Abstract

This study is aimed to develop an adaptation test to determine the adaptation level of the 6th grade students about the concepts and topics to their daily life included in the unit " Systems and Health in Our Body". The test items were prepared by considering the reasons of the problems that they can encounter in their daily life about the subject and the concepts. The test was reduced to 27 items depending on the opinions of the experts and the period that the students given to answer the items. The final application was conducted with 48 students after the necessary corrections were done. The evaluation of the test items was performed as an answer key and grading rubric were formed. In this study, Pearson correlation coefficient and the comparison of means theory was used to determine the scoring reliability among scorers.

Keywords: Science; controller and regulatory systems; endocrine system; sense organs; daily life

1. Introduction

To achieve the desired level of learning in the education process, different learning theories are put forward by the researchers. When current learning theories are examined, it is seen that the relation of course contents with daily life has an important place in internalizing information for the realization of learning (Aydın, 2000; Özden, 2003). Since it is desirable to transfer the information transferred during the teaching process, the association with daily life should have an important place in the process. It is stated by researchers that information that cannot be used in daily life does not go beyond memorization (Demircioğlu, Demircioğlu, Ayas, & Kongur, 2012; Ince, 2012).

Science is one of the branches of science that helps the individual to understand the environment in which he lives. Science can be defined as an effort to examine, understand, interpret and use the events, rules and laws in the environment (Isman, Baytekin, Balkan, Horzum & Kiyici, 2002). When science subjects are examined, it is

* Corresponding author name. Tel.: +09-5455681158
E-mail address: miyase.tutar@amasya.edu.tr

seen that individuals include situations that they may encounter in daily life. If science subjects and concepts are learned and used, it provides convenience in individuals' lives. Therefore, the importance of learning some basic information and using it in daily life is increasing day by day (Aktepe & Aktepe, 2008; Tekbıyık, 2015). It is as much as that people can use the information, they learn in their lives to benefit from these facilities (Gilbert, 2006).

Since science involves the situations encountered in daily life, it is necessary to look at the skills of students to apply what they have learned to daily life in order to determine whether they understand the subjects and concepts (Balkan-Kıyıcı & Aydoğdu, 2011; Gilbert, 2006; Hürcan-Gürler & Önder, 2014; Taşdemir & Demirbaş, 2010). It can be understood whether the aim of the teaching given can be achieved by considering the application of the subjects to daily life (Coştu, Ünal & Ayas, 2007; Gilbert, 2006).

To determine the level of students' association of science subjects with daily life, valid and reliable measurement tools appropriate to the level of students are needed. However, one of the biggest problems encountered in the assessment and evaluation processes is to develop an equal unit, generalizable and objective measurement tool that will help determine the extent to which behavior change occurs in accordance with the intended acquisitions (İlkörücü-Göçmençelebi and Özkan, 2010) Adapting the learned knowledge to daily life Since it requires skills, it is necessary to determine the appropriate measurement tools to be used in this process. It is seen that written exams are preferred especially when high level knowledge and skills of students are desired to be measured (Acar, 2018).

Written examinations are used to determine the original thoughts of the people, written expression skills are used and opinions about a topic are taken (Tekin, 2004). In written exams, students are asked to answer the question in their own sentences by asking open-ended questions. Written exams are the type of exams that the responder organizes and presents the written answers by remembering and remembering the answers to the questions asked (Atılgan, 2006). The reason why written exams differ from other types of exams is that the answers given require systematic and complex thinking. (Reiner, Bothell, Sudweeks and Wood, 2003). In addition, it allows students to measure more complex behaviors above the level of knowledge through written exams (Tekin, 2004).

Important factors that affect the quality of the written exams are the length of the answers and the necessity of giving written, the independence of the answers, the scoring is based on the rater's opinion. Measures should be taken to prevent these elements from adversely affecting the validity and reliability of the written exams (Atılgan 2006).

Analytical, classification, ranking and general impression scoring methods are used in the assessment of the written exams. The reliability of the scoring results of other methods except analytical scoring (key scoring) is low (Tekindal, 2000) In the scoring

with key method, answers to the questions are prepared in detail and scores are determined for each information. Since the scoring is based on the key determined at the beginning, this method is expected to give more reliable scores compared to other methods (Tekindal, 2000).

Rubric is a guide to determine the competence of the students' performance by using the answers and performances of the students according to the determined dimensions (Jonsson and Svingby, 2007; Turgut and Baykul, 2012). Teachers should determine the behaviors they want to measure in advance and prepare graded key according to these behaviors (Moskal, 2003).

In this study, it is aimed to develop a valid and reliable test to determine the students' ability to apply related science subjects in daily life. In order to apply science subjects in daily life, information needs to be transferred. Therefore, the application of science subjects in daily life requires high level skills. In the study, it was decided to use written tests, one of the most appropriate methods, to measure the existence of the mentioned high-level skills. In order to ensure the validity and reliability of the test, the section was prepared.

During the research process, literature review related to the application of science subjects to daily life was conducted. When the literature is examined, it has been found out that studies related to the concepts of biology, matter and heat, state change, force and motion and the systems in our body in relation to daily life were found in studies conducted with secondary school students (Emrahoğlu & Mengi, 2012; Er-Nas, 2008; Hürcan & Önder, 2012; İlkörücü-Göçmençelebi and Özkan, 2009; Şahin and Bodur 2016). In this research, the daily adaptation test was developed for the subjects and concepts in the Body and Systems Unit of the Body in the 2018 Science Curriculum. Since the test developed in the research process is prepared in accordance with the current curriculum and includes all the concepts within the unit, it is thought that it will contribute to the literature.

2. Method

In this research, a written and validated test was developed to determine the adaptation of the subjects and concepts in the "Systems and Health in Our Body" unit to the daily life of the 6th grade science class.

2.1. Participant (subject) characteristics and sampling procedures

The population of the study consists of 8th grade students studying in Amasya in the fall semester of 2018-2019, and the sample consists of 48 students randomly selected from this universe. As the test developed in the research, the 6th grade Science course was directed towards the 'Systems and Health in Our Body' unit and the 6th grade

students had not learned this unit during the course, the 8th grade students who learned the related subjects and concepts were selected and applied with these students.

2.2. Measures and covariates, data collection tool

In this study, a 27-item written test was developed for the cognitive area acquisition of the 'Systems and Health in Our Body' unit subjects and concepts as a data collection tool. The steps taken during the development of the written test are summarized below.

The steps of Güler's (2012) test development process were used in the preparation of the 'Adaptation to Daily Life Test'.

2.2.1. Determination of the purpose of the test

It was prepared with the test in order to determine the learning levels of the students according to the acquisitions in the 2018 Science Curriculum and to provide the practitioners with guidance about whether the unit and the concepts and concepts in the unit apply to daily life.

2.2.2. Determining the scope of the test

Within the scope of the test, the subjects, course hours, achievements and the levels and areas of achievements were determined according to the 2018 Science Curriculum. Detailed information in this unit is given in Tables 1 and 2.

Table 1. The information about Systems and Health in Our Body

Unit Name	Subject Area	Number of Acquisition	Course Hours
Systems and Health in Our Body	Living Creatures and Life	13	18

There are 11 acquisitions related to the "Systems and Health in Our Bodies" unit in the science curriculum. These acquisitions have been added to the acquisitions of 'Knowing system-based diseases! and 'Distinguishing organ donors' in line with the opinions of experts in the field education and the total number of acquisitions has been 13. The recommended time for teaching the unit is 18 hours.

2.2.3. Writing and correction of items

Prior to the preparation of the test items, different researches about reconciliation with daily life in the literature were examined. As a result of the researches, the items of the scale were decided to be open ended. It was preferred to use open-ended questions for the test as it provides the opportunity for students to express their knowledge and thoughts about the subject in detail (Atılgan, 2006). Before the writing of the items, the acquisitions in the science curriculum were analyzed and the levels of the acquisitions were determined.

While preparing the items, the situations that the students might encounter in daily life related to the acquisitions in the curriculum were taken into consideration. A total of 42 items were prepared by asking the reasons for the situations the students might encounter. The opinions of two field educators were taken during the preparation of the items and corrections were made in accordance with the feedback given. As a result of feedback and corrections, a test consisting of 27 items was prepared. Since the test was going to be applied during one lesson hour, questions that were time consuming to answer in the writing of articles were avoided. The item suitable for each acquisition is written and the test specification table is given in Table 2.

Table 2. Specification Table

Nervous System	**Comprehension**	**Application**	**Analysis**	**Synthesis**	**Total**
Explains the nervous system, central and peripheral nervous system functions on the model.		1,2	3		3
Realize the importance of internal secretion glands for the body. (Describes)	4	5,6			3
Explain the physical and mental changes that occur during the transition from childhood to adolescence.	7,8				2
Discusses what can be done in order to spend adolescence in a healthy way based on research data			9		1
Discuss the effect of supervisory and regulatory systems on the regular and coordinated operation of other systems in our bodies.	10				1
Sense Organs					
Explain the structures of sense organs by showing them on the model.	12	13,15	11	14	5
It shows the relationship between the senses of smell and taste, with an experiment designed		16			1
It gives examples of defects in sense organs and the technologies used to eliminate these defects.	17				1
Discuss the measures to be taken to protect the health of sense organs.	18,19	20,21,22			5
The Health of Systems					
Knows the diseases related to systems.		24,26,27			3
Discusses what needs to be done for the health of systems based on research data.	23	25			2
Comprehend the importance of organ donation in terms of social solidarity.					
Distinguishes those who can donate organs.					
Total	9	14	3	1	27

When Table 2 is examined, it is seen that more items are written for some acquisitions than others. Due to the high content of the related acquisitions, the number of substances measuring these acquisitions is higher than the others. In addition, the absence of any

article on organ donation is since the 8th grade students in whom the application was made have not yet learned about this subject.

2.2.4. Validity of the test

For the content and appearance validity of the written test, 8 expert opinions were obtained from 2 faculty members in science education, one faculty member in measurement and evaluation, and 5 science teachers. For the examination of the test in terms of grammar, one Turkish field education expert was consulted. As a result of the suggestions and corrections from the expert opinions, the final version of the test was established.

2.2.5. Pre-pilot application

Before the pilot application, the pre-pilot application was conducted with 18 8th grade students in order to determine whether the questions in the written test were understood by the students and whether the time was sufficient during the implementation of the questions and they learned the subject. It was seen that the test can be answered in one lesson (40min) and the questions are understandable by the students.

2.2.6. Pilot application

The pilot application of the test was conducted with 8th grade students studying in the fall semester of 2018-2019 in Amasya. The reason for the selection of 8th grade students in the pilot application is that these students have seen the subjects and concepts in the achievement test in the previous periods. The achievement test was applied to 48 8th grade students and the analysis of the test was done with the data obtained from 48 students.

2.2.7. Item analysis

Before the analysis of the test answers, the categories of Understanding (U), Partial Understanding (PU), Misunderstanding (MA) and Blank-Unanswered (B) were used and the answer key and scoring rubric were created accordingly. These categories are frequently used in the literature (Abraham, Grzybowski, Rennerve Marek, 1992; Özmen, 2003). Table 3 shows the categories and scores of the assessment rubric.

Table 3. The Assessment rubric.

Student Answer	Category	Point
Explain matter with correct concepts	Understanding (U)	5
Use only the right concept or explain why without using the concept	Partial Understanding (PU)	2,5
Incorrect answer	Misunderstand (MU)	0
Nonresponse	Blank (B)	0

A rubric was prepared to examine the scorers during the scoring. The process of preparing a rubric was carried out in accordance with the steps proposed by Goodrich (2001):

1. Determination of criteria lists: Criteria appropriate to the purpose of the test were determined.

2. Determination of the type of section: For the purpose of the test, analytical Rubric was used.

3. Determination of performance levels: The responses to the test items are written in detail.

4. Obtaining expert opinions about the section: After the section was formed, the necessary corrections were made by taking the opinions of 1 biology education specialist and 1 science teacher.

The answers to the test were scored by two different raters (one science education expert and one science teacher). The scorers scored the answers of the students according to the sections prepared independently from each other. After the scoring process, item analyzes were performed.

In the written tests, item analyzes are performed in different ways than the other tests. Since open-ended items were used in the written tests, the item difficulty index (p_j) was calculated by taking the arithmetic average of the scores obtained by the respondents to the highest score determined for that item (Acar, 2018). Item discrimination power index (r_{jx}) is calculated by the correlation coefficient of the product and test scores because of the continuous variable pearson moments product (Baykul, 2000).

2.3. Data Analysis

Different methods are used to determine the reliability of the scores of the raters in the written tests (Güler and Taşdelen Teker, 2015). In this study, Pearson correlation coefficient and the comparison of means theory was used to determine the scoring reliability among scorers. SPSS program was used for the calculation of Pearson correlation coefficient and comparison of means. The Pearson correlation coefficient and the comparison of the means were calculated separately for both the whole test and for each item.

2.3.1. Pearson correlation coefficient

The Pearson correlation coefficient, a frequently used method for calculating inter-rater reliability, is defined as the consistency of the scores of the two scorers (Baykul, 2000).

2.3.2. Comparison of means

The correlation value between scorers' scores is independent of the mean (Goodwin and Goodwin, 1991). The high correlation between the scores of the scorers does not indicate that the averages are close to each other. Therefore, the calculation of the correlation coefficient alone is not expected to fully reflect inter-rater reliability. For this reason, to determine the reliability between scorers, t-test was used to look at the differences between the means (Güler & Taşdelen Teker, 2015).

3. Results

3.1. Statistics and data analysis

This section contains difficulty levels of items included in the test, Pearson correlation coefficient values for inter-scorer reliability and t-test results are comparison of means.

Difficulty levels of items are given in Table 4.

Table 4. Difficulty levels of items (p_j)

Item Number	P_j values by scorer 1	Difficulty levels	P_j values by scorer 2	Difficulty levels
1	0,32	intermediate difficulty	0,34	intermediate difficulty
2	0,23	too difficulty	0,25	too difficulty
3	0,06	too difficulty	0,05	too difficulty
4	0,21	too difficulty	0,20	too difficulty
5	0,36	intermediate difficulty	0,38	intermediate difficulty
6	0,25	too difficulty	0,21	too difficulty
7	0,13	too difficulty	0,11	too difficulty
8	0,09	too difficulty	0,07	too difficulty
9	0,35	intermediate difficulty	0,33	intermediate difficulty
10	0,09	too difficulty	0,09	too difficulty
11	0,40	intermediate difficulty	0,41	intermediate difficulty
12	0,44	intermediate difficulty	0,44	intermediate difficulty
13	0,63	easy	0,63	easy
14	0,03	too difficulty	0,03	too difficulty
15	0,29	too difficulty	0,29	too difficulty
16	0,41	intermediate difficulty	0,41	intermediate difficulty
17	0,57	easy	0,57	easy
18	0,50	easy	0,50	easy
19	0,33	intermediate difficulty	0,33	intermediate difficulty

20	0,17	too difficulty	0,19	too difficulty
21	0,17	too difficulty	0,17	too difficulty
22	0,18	too difficulty	0,15	too difficulty
23	0,33	intermediate difficulty	0,34	intermediate difficulty
24	0,08	too difficulty	0,08	too difficulty
25	0,11	too difficulty	0,11	too difficulty
26	0,13	too difficulty	0,16	too difficulty
27	0,45	intermediate difficulty	0,43	intermediate difficulty
Test	**0,27**	too difficulty	**0,27**	too difficulty

When the Table 4 is examined, it is seen that there are easy, difficult and intermediate difficult items in the test (Turgut & Baykul, 2012). In addition, difficulty levels of items were at the same levels according to both scorers. When the whole test is examined, it is seen that the test is difficult for the students. When the table is examined, it is seen that the most difficult items for the students are 3, 8, 10 and 11. These items are listed below:

- Why are our hands numb when we are in a very cold environment for a long time?
- What is the reason of adolescents' being often rebellious?
- What makes a paralyzed person unable to walk?
- Why do our eyes bleed when we are tired or sleepless?

The most easily answered items were items 13, 17 and 18. These items are listed below:

- Why do we perceive the taste of the bitter nutrients we eat?
- What is the reason for voice information on municipal buses and traffic lights?
- Why is it harmful to listen to music at high volume?

3.1.1. Pearson correlation coefficient

The answers obtained from 48 students were scored by 2 different raters considering rubrics. Pearson correlation coefficient was calculated for each item and overall test.

Table 5. Basic statistics according to scorers

Scorer	Mean	Standard Deviation	Pearson correlation coefficient
Scorer 1	36,7188	14,83200	,985
Scorer 2	36,4583	14,49352	

In Table 5, mean, standard deviation and Pearson correlation coefficient between scorers are given. When the table is examined, it is seen that the scorers mean, and standard deviation values are close to each other. In addition, the Pearson correlation coefficient between raters is ,985.

Table 6. Pearson correlation coefficient of items between scorers

Item Number	Pearson correlation coefficient	Item Number	Pearson correlation coefficient
1	,960	**15**	,961
2	,959	**16**	,968
3	,902	**17**	1,000
4	,964	**18**	1,000
5	,955	**19**	,906
6	,879	**20**	,893
7	,830	**21**	1,000
8	,709	**22**	,903
9	,893	**23**	,961
10	1,000	**24**	,850
11	,944	**25**	1,000
12	1,000	**26**	,902
13	1,000	**27**	,920
14	1,000	**Test**	,985

Pearson correlation coefficient between raters for each item are given in Table 6. When Table 6 is examined, it is seen that the lowest correlation value is ,830 (7) and the highest correlation value is 1,000 (10, 12, 13, 14, 17, 18, 21 and 25). Pearson correlation coefficient is if the relationship is greater than 0.70 and the relationship is high (Büyüköztürk, Çokluk Bökeoğlu & Köklü, 2009).

3.1.2. Comparison of means

Another approach that can be used to analyze inter-rater reliability is to compare the means. Table 2 presents the results of the comparison of the differences between the raters' scores. t-test was performed between the mean scores of the raters for each item and the test.

Table 7. t-test results between scorers

Item Number	Average of the scorer 1	Average of the scorer 2	Difference Between Means	t	p
1	1,6146	1,7188	-,10417	-,294	,769
2	1,1979	1,2500	-,05208	-,202	,840
3	,3125	,2604	,05208	,317	,752
4	1,0417	,9896	,05208	,190	,850
5	1,8229	1,8750	-,05208	-,219	,827
6	1,2500	1,0417	,20833	,727	,469
7	,6250	,5729	,05208	,237	,813
8	,4688	,3646	,10417	,543	,589
9	1,7708	1,6667	,10417	,327	,744

10	,4688	,4688	,00000	,000	1,000
11	1,9792	2,0313	-,05208	-,136	,892
12	2,1875	2,1875	,00000	,000	1,000
13	3,1250	3,1250	,00000	,000	1,000
14	,1563	,1563	,00000	,000	1,000
15	1,4583	1,4583	,00000	,000	1,000
16	2,0313	2,0313	,00000	,000	1,000
17	2,8646	2,8646	,00000	,000	1,000
18	2,5000	2,5000	,00000	,000	1,000
19	1,6667	1,6667	,00000	,000	1,000
20	,8333	,9375	-,10417	-,330	,742
21	,8333	,8333	,00000	,000	1,000
22	,8854	,7292	,15625	,552	,582
23	1,6667	1,7188	,05208	-,198	,843
24	,4167	,4167	,00000	,000	1,000
25	,5729	,5729	,00000	,000	1,000
26	,6250	,6771	-,05208	-,179	,858
27	2,2396	2,1354	,10417	,418	,677
Test	**1,3561**	**1,3426**	**,01350**	**,122**	**,903**

*p<0,05

In table 7, results of t test between scorers means of the scores for the items and the test are given.

When Table 7 is examined, it is seen that there is no significant difference between the scorers regarding the mean of the items and the whole test (.05<p).

4. Discussion and Conclusions

After the analysis of the 27-item test developed, no item was removed from the test. It was found that more than half of the items in the test were difficult. In addition, the test of adapting to daily life was a difficult test for the students. It is thought that the students may have forgotten the concepts contained in the test in the previous years as the reasons for the small number of answers of the test items, or that the concepts related to the daily life could not be given in the teaching process. One of the reasons for the low average of the test may be that students are willing to answer questions that require open-ended answers and are reluctant to write the answer in detail.

Correlation coefficient values are examined for the value of discrimination in open-ended items (Özçelik, 1997: 123). In this case, it is desirable that the correlation coefficients of the items in the test take values close to 1.00. Because an item with a value close to 1.00 distinguishes between having and no-having the measured property (Acar, 2018). When Pearson correlation coefficients of test items are examined, it is seen

that the lowest value is 0,709 and the highest value is 1,000. By looking at the values, it can be said that the test items are distinctive.

To determine the reliability of the test prepared during the research process, the reliability of the raters be determined. The Pearson correlation coefficient and comparison of means were used to determine the consistency between the scores of the scorers.

A Pearson correlation coefficient between 0.00 and +1.00 indicates a positive relationship. As the correlation coefficient approaches +1.00, it shows that there is a strong relationship between the two variables in the same direction (Büyüköztürk, 2012). In this study, the correlation value calculated between raters was found to be 0.985. When this value is analyzed, it is seen that there is a high level and positive relationship between the scorers' scores.

The correlation coefficients obtained in the reliability calculations of the written tests do not show similarities and differences between the raters. Therefore, the calculation of the correlation coefficient may be insufficient in determining the inter-rater compliance (Goodwin, 2001). Therefore, the reliability of the test was tried to be determined by looking at the differences between the means.

To determine the significance level of the differences between the scores obtained from the scorers and the total score averages of each item separately and for each item, t-test was performed. When the results of t-test between raters were examined, it was found that there was no significant difference between the averages both for the test and for all items ($0.05<p$). This shows that there is no significant difference between the scorers' averages, and that the test has rater reliability (Güler & Taşdelen Teker, 2015).

The use of multiple methods together in determining the reliability of the rater is important to achieve realistic results for the current situation and for less margin of error. Using a variety of methods to look at the reliability of the tests from a different perspective will help the test to be effective in the measurement.

In this study, the reason for the high correlation coefficient between raters and the fact that the rater averages are very close to each other is thought to be the preparation of the section before the evaluation. Since the analytical section clearly explained what points the two raters would give and what points would not be given, the raters' scores were close to each other. The use of the section is important to ensure the rater reliability is at the desired level (Kutlu, Doğan & Karakaya, 2014; Tekindal 2000).

References

Abraham, M. R., Grzybowski, E. B., Renner, J. W., & Marek, E. A. (1992). Understandings and misunderstandings of eight graders of five chemistry concepts found in textbooks. *Journal of Research in Science Teaching, 29*(2), 105-120. https://doi.org/10.1002/tea.3660290203

Acar, T. (2018). The impact of scoring method in written examinations on test and item statistics. *Elementary Education Online*, *17*(2), 500-509. doi 10.17051/ilkonline.2018.418890

Aktepe, V. & Aktepe, L. (2009). Teaching method using science and technology education on students' aspects: the example of Kırşehir, *Journal of Kırsehir Education Faculty*, *10*(1), 69-80.

Atılgan, H. (Ed.). 2006. *Eğitimde Ölçme ve Değerlendirme*, Ankara: Anı Yayıncılık.

Aydın, A., (2000). *Gelişim ve Öğrenme Psikolojisi.* İstanbul: Alfa Yayınları.

Balcı, A. (2001) *Sosyal Bilimlerde Araştırma: Yöntem, Yeti ve İlkeler,* Ankara: Pegem Yayıncılık.

Balkan-Kıyıcı, F. & Aydoğdu, M. (2011). Determination of pre-service science teachers' levels of relating the scientific knowledge to their daily lives. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 5*(1) 43-61,

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

Büyüköztürk, Ş. (2014). *Sosyal Bilimler İçin Veri Analizi El Kitabı. (19. Baskı*). Ankara: Pegem A Yayıncılık.

Büyüköztürk, Ş., Çokluk, Ö. & Köklü, N. (2010). *Sosyal Bilimler İçin İstatistik (5. Baskı).* Ankara: Pegem Akademi.

Coştu, B., Ünal, S. & Ayas, A. (2007). The use of daily-life events in science teaching. *Journal of Kirsehir Education Faculty, 8*(1), 197-207.

Demircioğlu, H., Demircioğlu, G., Ayas, A. & Kongur, S. (2012). A comparison of theoretical and practical knowledge about physical and chemical changes the concept of tenth grade students, *Journal of Turkish Science Education 9*(1), 162-181.

Emrahoğlu, N. & Mengi, F. (2012). An investigation of 8th grade primary school students' level of transferring science and technology topics into everyday life problem solving. *Journal of Çukurova University Institute of Social Sciences*, *21*(1), 213-228.

Er-Nas, S. (2008). Determining effectiveness of the materials about "the ways of spread of heat" based on the elaborate stage of the 5E model, Master Thesis, K.T.U., Institute of Science, Trabzon.

Gilbert, J. K. (2006) On the nature of "context" in chemical education. *International Journal of Science Education*, *28* (9), 957-976. https://doi.org/10.1080/09500690600702470

Güler, N. (2012). *Eğitimde Ölçme ve Değerlendirme. (2. Baskı)* .Ankara: Pegem Akademi.

Güler, N. & Taşdelen Teker, G. (2015). The evaluation of rater reliability of open-ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology, 6* (1), 2-24. DOI: 10.21031/epod.63041

Goodwin, L. D. & Goodwin, W. L. (1991). Using generalizability theory in early childhood special education. *Journal of Early Intervention*, 193-204. https://doi.org/10.1177/105381519101500208

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-14. https://doi.org/10.1207/S15327841MPEE0501_2

Ilkörücü-Göçmençelebi, Ş. & Özkan, M. (2009). Effect of elementary school sixth grade students' levels of connecting biological knowledge with everyday life on their achievement, *Kastamonu Education Journal, 17*(2), 525-530.

Ilkörücü-Göçmençelebi, Ş. & Özkan, M. (2010). A study of scale development testing elementary school 6th grade students' level of relating the knowledge given in science curriculum with everyday life. *Journal of Uludağ University of Faculty of Education, 23*(1), 121-132.

Ince, E. (2012). Investigation of undergraduate students' skills to associate between general physics concepts and everyday life. *International Journal of Academic Research Part B*, *4*(6), 158-163.

Işman, A., Baytekin Ç., Balkan, F., Horzum, M.B. & Kıyıcı M. (2002). Fen bilgisi eğitimi ve yapısalcı yaklaşım. *The Turkish Online Journal of Educational Technology, 1*(1), 1303-6521,

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kutlu, Ö., Doğan, C. D. & Karakaya, İ. (2014). *Ölçme ve Değerlendirme Performansa ve Portfolyoya Dayalı Durum Belirleme.* Ankara: Pegem Akademi.

Moskal, B. M. (2000). Scoring rubrics: what, when and how? *Practical Assessment Research and Evaluation, 7*(3), 1-5.

Özden, Y. (2003). *Öğrenme ve Öğretme*. Ankara: Pegem Yayıncılık.

Özçelik, A. D. (1997). *Test Hazırlama Kılavuzu.* Ankara: ÖSYM yayınları

Özmen, H. (2003). Chemistry student teachers' levels of linking their knowledge with daily life about acid and base concepts. *Kastamonu Education Journal, 11*(2), 317-324.

Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2003). *Preparing Effective Essay Questions*, Oklahoma: New Forums Press.

Şahin, Ç. & Bodur, Ş. (2016). Developing of the association test with daily life elementary school 7th grade 'systems in our body' unit. *Elementary Education Online*, *15*(3), 999-1016. doi: http://dx.doi.org/10.17051/io.2016.85699

Taşdemir, A. & Demirbaş, M. (2010). The level of correlation of concepts that primary students seen topics in science and technology class with daily life. *Journal of Human Sciences, 7*(1), 124-148.

Tekbıyık, A. (2015). The real-life application of pulleys in a competitive environment. *Teaching Science*, *61*(1), 18-26.

Tekin, H. (2004). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Yargı Yayınevi.

Tekindal, S. (2000). Klasik yazılı sınavla ve çok sorulu testle elde edilen ölçümlerin güvenirlik ve geçerliği. *Pamukkale University Journol of Education*, *8*(8), 38–46.

Turgut, M. F. & Baykul, Y. (2012). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Pegem Akademi.