Examination of Misconceptions in the Field of Alternative Measurement and Evaluation with A Four-Tier Test

Öğr. Gör. Tansu ALAN*

Adıyaman Üniversitesi, Rektörlük, Eğitimde Ölçme ve Değerlendirme, Adıyaman / Türkiye,
tansualan@adiyaman.edu.tr, ORCID: 0000-0001-5855-0302

Doç. Dr. Ufuk AKBAŞ

Hasan Kalyoncu Üniversitesi, Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Bölümü,
Gaziantep / Türkiye, ufuk.akbas@hku.edu.tr, ORCID: 0000-0002-6122-154X

Abstract

The aim of this study is to examine teachers' misconceptions in the field of alternative measurement and evaluation with a four-tier test. For this purpose, a four-tier misconception test consisting of 15 items has been developed. Test-retest and KR-20 method have been used for the reliability of the misconception test developed in the study. For the validity of the test, expert opinions have been taken, false positive and false negative percentages have been calculated, the relations between the scores obtained from the different tiers of the test have been examined, and the relationship between the scores obtained from a two-tier test and the test developed in this study has been examined. The sample of the research consists of 360 teachers working in schools at different levels in Adıyaman in the 2020-2021 academic year. In the analysis of the

* Sorumlu Yazar. Tel: +90 416 223 38 00 / 3312 | Araştırma Makalesi.

Makale Tarih Bilgisi. Gönderim: 20.01.2023, Kabûl: 15.12.2023, Erken Görünüm: Ağustos, 2024, Basım: Haziran, 2025

data, besides the statistics such as frequency percentage, test and item statistics and correlation analysis has been used. It has been determined that teachers generally have misconceptions and lack of knowledge, albeit partial, on alternative measurement and evaluation issues. Teachers mostly have misconceptions about the performance task, and they have lack of knowledge mostly about the structural grid.

Keywords: Alternative measurement and evaluation; Four-tier test; Misconception; Validity; Reliability.

Tamamlayıcı Ölçme ve Değerlendirme Alanındaki Kavram Yanılgılarının Dört Aşamalı Testle İncelenmesi

Öz

Bu çalışmanın amacı öğretmenlerin tamamlayıcı ölçme ve değerlendirme alanındaki kavram yanılgılarının dört aşamalı test ile incelenmesidir. Bu amaç doğrultusunda 15 maddeden oluşan dört aşamalı kavram yanılgısı testi geliştirilmiştir. Çalışmada geliştirilen kavram yanılgısı testinin güvenirliği için test tekrar test ve KR-20 yöntemi kullanılmıştır. Testin geçerliği için uzman görüşleri alınmış, pozitif yanlış ve negatif yanlış yüzdelikleri hesaplanmış, testin farklı aşamalarından alınan puanlar arasındaki ilişkilere bakılmış ve iki aşamalı bir test ile bu araştırmada geliştirilen testten alınan puanlar arasındaki ilişkiye bakılmıştır. Araştırmanın örneklemini, 2020-2021 eğitim öğretim yılında Adıyaman ilinde farklı kademedeki okullarda görev yapan 360 öğretmen oluşturmaktadır. Verilerin analizinde frekans yüzde gibi istatistiklerin yanında test ve madde istatistikleri ile korelasyon analizinden yararlanılmıştır. Öğretmenlerin genel olarak tamamlayıcı ölçme ve değerlendirme konularında kısmi de olsa kavram yanılgısı ve bilgi eksikliği yaşadığı belirlenmiştir. Öğretmenler en çok performans görevi konusunda kavram yanılgısı yaşamakta, en çok yapılandırılmış grid konusunda bilgi eksikliği yaşamaktadırlar.

Anahtar kelimeler: Tamamlayıcı ölçme ve değerlendirme; Dört aşamalı test; Kavram yanılgısı; Geçerlik; Güvenirlik.

Introduction

Schools exist to train the members of a society and to prepare them for a successful future. However, traditional education practices cannot handle this process and the graduates cannot become suitable for the needs of contemporary societies (Ün-Açıkgöz, 2003). Today, in addition to having just knowledge, individuals are expected to have skills such as finding ways to access knowledge, transferring knowledge to different environments and situations, building on existing knowledge, analyzing and synthesizing knowledge, scientific, critical and creative thinking, cooperation, and good communication. With the 2005 curriculum, which was prepared to meet the expectations of today, the behavioral approach, which focused on behaviors for many years, was abandoned and the constructivist approach, in which the student constructs knowledge, was embraced. The use of alternative measurement tools and directions for these tools were included in the 2005 Turkish curriculum (Özmantar et al., 2018). Alternative measurement and evaluation methods include all methods that are outside of traditional measurement and evaluation methods, such as multiple-choice test and true-false methods (Bahar et al., 2015). These methods include such as portfolio, rubric, performance assessment, observation, project, self-assessment, peer assessment, group assessment, interview, branching tree, word association test, structured grid, feedback, authentic assessment (Akbaş et al., 2018; Bahar et al., 2015). The common feature of these methods is that the student is active in the learning process. These methods, which are effective in measuring and assessing higher-order behaviors, can measure skills such as critical and creative thinking, the correct use of scientific concepts, establishing a connection between fundamental concepts and daily life, appropriate use of sources and references, and the ability to synthesize information and ideas (Bekiroğlu, 2004). For instance, students can actively participate in education through self-assessment, peer assessment, and group assessment, leading to an increase in interest, motivation, communication, and critical thinking skills, as well as improved academic achievement levels (Kutlu et al., 2014). Alternative measurement and evaluation methods have become more important in the education process subsequent to the new changing Turkish curriculum. In the alternative assessment and evaluation activities, the student is active in the process and can access to the knowledge on his/her own instead of memorizing the knowledge discerned from the teacher and can use the knowledge s/he has received in different situations. One of the most important features of these methods is that they activate high-level mental, affective and psycho-motor skills (MEB, 2005; 2009). Therefore, not only it is important to learn these methods correctly but also it is important to use these methods in education. If the alternative measurement and evaluation methods, which include many concepts, are not structured correctly, mistakes may arise. For example, an alternative measurement and evaluation method applied without knowing the purpose of it may lead to unscientific interpretations and may not serve its purpose. These unscientific interpretations lead to misconceptions.

An individual learns various concepts as a result of his/her own life and experiences, communication and interaction with his/her environment. The individual encounters these concepts at different stages of the education-teaching process. If the concept that takes place in the mind of the individual is defined correctly, scientific learning happens easily. If the concepts are shaped in the mind of the student in a way that is far from being scientific, it is very difficult to destroy these misconstrued concepts. These misconstrued concepts are called misconceptions. Misconceptions are scientifically incorrect thoughts (Leonard et al., 2014). Misconceptions include understanding or thinking which is not based on true information. Misconceptions occur because of errors in transferring concepts from information obtained into a framework. So, the concept understood may not be in accordance with the actual concept (Burgoon et al., 2017). A misconception is not a wrong answer given by the student due to an accidental mistake or lack of knowledge. The misconception is that the concept in the mind of the individual is far from the scientific definition. If the individual explains the accuracy of his/her mistake by giving reasons and expresses that s/he is sure in these explanations, then it can be said that there is a misconception (Eryılmaz and Sürmeli, 2002). Since misconceptions are very resistant to change and can create problems for more scientific knowledge, it is very important to identify misconceptions (Kaltakçı-Gürel et al., 2015; Smith et al., 1994). Many misconception tests can be developed for researchers to use in the detection of misconceptions, and these misconceptions can be reached by conducting one-on-one interviews (Güneş, 2005). When the literature is examined, there are many different methods used by researchers in determining misconceptions (Caleon and Subramaniam, 2010a; Fratiwi et al., 2017; Kaltakçı, 2012; Kanlı, 2015; Karadeniz-Bayrak, 2013; Lin et al., 2015; Milenkovic et al., 2016; Peşman and Eryılmaz, 2010; Treagust, 1988). These methods are multiple choice tests, two-tier, three-tier and four-tier tests and these methods have various advantages and disadvantages compared to each other.

The fact that they can be applied to a large group and the results can be easily analyzed has led researchers to use multiple choice tests. However, considering the definition of the misconception, it limits the use of these tests due to the inability to distinguish it from error and lack of knowledge. (Eryılmaz and Sürmeli, 2002). Since there is a chance factor in these tests, it is very difficult to predict whether the student reached the right answer on purpose or by chance. Another missing point is whether these tests reveal lack of knowledge or misconceptions. Since these tests do not have a test system that will reveal the reason for choosing the option chosen by the student, two-tier, three-tier and four-tier tests have been developed (Caleon and Subramaniam, 2010b).

In the two-tier test, the first tier is defined as a diagnostic test consisting of multiple-choice questions. The second tier consists of options including the explanations of the answer given in the first tier (Kaltakçı-Gürel et al., 2015; Treagust, 1986). These tests cannot distinguish whether the error is due to a lack of knowledge or a misconception. At the same time, it cannot be decided whether the correct answer was reached intentionally or by guess (Bagayoko and Keller, 1999; Caleon and Subramaniam, 2010; Hasan et al., 1997). Due to this weakness of the two-tier tests, misconceptions can be significantly addressed by adding a third step, called the reliability level, which measures the reliability of the participants in the answers given in the first two tiers (Caleon and Subramaniam, 2010a). Although these three-tier tests are thought to be a way to measure misconceptions independently from errors and lack of knowledge, there are still some limitations due to the latent grading of reliability in the first and second tiers of these tests. This issue can cause two problems: the first is the underestimation of the lack of knowledge rate, and the second is the overestimation of students' misconceptions and correct answers (Kaltakçı, 2012). Due to the limitations of the single reliability level, four-tier tests have been developed to measure reliability in both tiers. Thus, reliability in both tiers has been measured in separate tiers. While the four-tier tests preserve all the strengths provided by the three-tier tests, they truly evaluate the misconceptions regardless of lack of knowledge and error (Kaltakçı-Gürel et al., 2015). Although it is known that the four-tier tests give more accurate results than other tests in identifying misconceptions, these tests have some limitations such as requiring a very long time and false reactions resulting from social likability (Caleon and Subramaniam, 2010a; Caleon and Subramaniam, 2010b). Due to the advantages of four-tier tests over other tests, it is aimed to reveal misconceptions with the help of a four-tier test in this study. For this purpose, "What is the level of teachers' misconceptions, scientific knowledge and lack of knowledge in the field of alternative assessment and evaluation?'' question has been tried to responded.

When the literature is examined, it is seen that almost all of the researches carried out to identify misconceptions with the help of four-tier tests are carried out in fields such as physics, chemistry and biology (Caleon and Subramaniam, 2010a; Görkemli-Taban, 2017; Eryılmaz and Sürmeli, 2002; Kaltakçı, 2012; Kılınç, 2017; Meşin, 2019; Önsal, 2012; Sheppard, 2006; Smith et al., 1994; Sreenivasulu and Subramaniam, 2013; Yang, 2019). There are fewer studies to identify misconceptions in the field of measurement and evaluation (Arık, 2006; Demirbilek, 2015; Üztemur, 2013). Arık (2006) identified teachers' misconceptions in the field of measurement and evaluation with a two-tier misconception test he developed. According to the findings of the study, it was concluded that the most misconceptions of teachers were in the concept of "correct scoring" with a rate of 40%. A similar study was conducted by Üztemur (2013). The findings were similar to Arık (2006) and it was determined that there was a misconception in the concept of "correct scoring" with a rate of 43.4%. Demirbilek (2015) determined the misconceptions of pre-service teachers in the field of measurement and evaluation with a two-tier misconception test she developed. In the research, it was seen that the pre-service teachers' mostly made common mistakes in the concepts of "difficulty index" and "normal distribution". When examining research conducted outside of measurement and evaluation, it has been often observed that four-tier tests are predominantly used to investigate misconceptions. With the help of these tests, individuals' levels of misconception and lack of knowledge have been determined. In these studies, it has been stated that four-tier tests are more reliable than two and three-tier tests in identifying misconceptions (Caleon and Subramaniam, 2010b; Fratiwi et al., 2017; Önsal, 2016; Sreenivasulu and Subramaniam, 2013).

The number of studies that will reveal the mistakes of teachers in measurement and evaluation in education is relatively less compared to other fields. Accordingly, no previous research has been found to investigate the misconceptions in alternative measurement and evaluation, which is the subject of this research. Teachers who lack knowledge and misconceptions about alternative measurement and evaluation methods can not be able to accurately assess their students' development and effectively guide the educational processes. Therefore, it is important for teachers to be informed about alternative measurement methods and to be able to use these methods effectively in order to improve the quality of education. Identifying teachers' misconceptions and lack of knowledge in this regard can be a step towards improvements in the field of education. Therefore, examining this subject has been thought to be important. For this purpose, this research tries to determine the teachers' misconceptions, lack of knowledge and scientific knowledge about alternative measurement and evaluation methods with a four-tier test.

Method

This section details the methodology used in this study. Research method, population and sample, development process of the measurement tool, data collection and coding and analysis of data are presented in this section.

Research Method

In this study, the survey model, which is among the quantitative research designs, has been used. Survey model is used to collect and analyze data in order to reveal certain characteristics of a group (Büyüköztürk et al., 2018). Ethics committee approval of this study was obtained with the decision no E--804.01-BABBFCF3 of Hasan Kalyoncu University Social and Human Sciences Ethics Committee at the meeting dated 03.11.2020.

Population and Sample

The population of this study consists of teachers working in state and foundation schools in Turkey. The sample of the research consists of teachers working in public and foundation schools in Adıyaman. Accordingly, the sample consists of 360 teachers in total. The convenience sampling method, one of the non-random sampling methods, has been used to determine the sample. In the study group, according to the branches, Elementary School (19.7%) (n=71), Primary School Mathematics (10.3%) (n=37), and English (7.5%) (n=27) Teachers have the highest rates. The average seniority of the teachers whose seniority ranges from 1 to 33 is 12.3 (SS=7.9). The distribution of teachers in the sample by gender and school level is given in Table 1 below.

Table 1. Distribution of Teachers in the Study Group by Gender and School Level

		School Level			**Total**	**Percentage**
		Primary school	**Secondary School**	**High School**		
Gender	F	46	81	44	171	47.5
	M	41	69	79	189	52.5
Total		87	150	123	360	100
Percentage		24.2	41.7	34.2	100	

F(Female), M(Male)

Development Process of the Measurement Tool

While developing the Alternative Measurement and Evaluation Misconception Identification Test (AMEMIT), Treagust's (1986, 1988) two-tier misconception test development tiers have been used. In the test development process, besides the literature review, the most recently published 2017 curriculum has been examined. With the 14 alternative measurement and evaluation methods in the curriculum, the concept of alternative measurement and evaluation has been discussed (Bahar et al., 2015; MEB, 2017). In order to collect data about the teachers' prior knowledge about the determined concepts, their misconceptions and their wrong and lack of learning, firstly, open-ended items and multiple-choice in the first part and open-ended items in the second part have been written about each concept. After examining the data collected from the pre-service teachers, the items for the test have been written. The first and third tiers of the test are in multiple-choice item format. The first tier is an achievement test in which the knowledge of the teachers is measured, and the third one is the tier in which the reason for the answer chosen in the first tier is demanded. The second and fourth-tiers are the same in which reliability is measured by choosing sure-not sure boxes. Two items have been written for each concept in the test and the test has been completed with a total of thirty items. The four-tier misconception identification test has been presented to the opinion of four experts working. Below, you can see the a four tier item in the test.

1. If Kübra Teacher wants to see the progress of her students during the teaching process and to include them in the evaluation process, which of the following is the most appropriate measurement and evaluation method that she can use?
   A. Self-assessment
   B. Authentic Assessment
   C. Performance Assessment
   D. Portfolio Assessment

1. 2. Are you sure about your answer to the above question?

□ Sure □Not Sure

1. 3. Because

A. The scoring is objective.
B. It is more reliable than objective tests.
C. It is good at measuring the knowledge and comprehension step.
D. Provides permanent learning.

1.4. Are you sure about the justification you chose above?

□ Sure □Not Sure

As a result of expert opinions, final corrections have been made to the test and it has been made ready for the pilot scheme. Pilot scheme data has been collected through the Google form due to the Covid-19. The pilot scheme has been made with 100 teachers. Difficulty and discrimination have been calculated in MS Excel for both the first tier of the test and the second tier including justifications. The pre-study has been made with 8 teachers and the full-scale study has started. As a result of the full-scale study, the forms with missing markings have been removed and analyzes have been made with 360 teachers who have made complete markings. Data have been collected for test-retest with one of the two groups different from the full-scale study, and for criterion validity with the other group. There are 54 teachers in the test-retest group and 51 teachers in the criterion validity group.

As a result of the item analyzes made with the data obtained as a result of the pilot scheme, 14 items are selected with item difficulties between 0.24 and 0.70 for both the first and the third tier and with item discrimination between 0.26 and 0.78. One item has been corrected and put to the test. The average difficulty for the first tier of the test is 0.45, and the average discrimination is 0.40. The average difficulty and discrimination for the third tier, which includes the reasons for the answers given, is 0.48.

To determine the reliability of the test, test-retest and KR-20 reliability have been examined. For test-retest reliability, the test has been administered to 54 pre-service teachers 15-20 days, and the reliability is calculated as 0.74 for the first tier of the test and 0.78 for the third tier. These reliability results reveals that the test scores have not changed much in the two applications and are stable. KR-20 reliability is examined to determine the internal consistency of the test. KR-20 reliability is 0.49 for the first tier of the test, and 0.57 for the third tier of the test. According to Salvucci et al. (1997) less than 0.50, the reliability is low, between 0.50 and 0.80 the reliability is moderate and greater than 0.80, the reliability is high. This criterias show that the first tier of the test is low and the third tier of the test is moderate. While the test-retest reliability coefficients are at an acceptable level, the low internal consistency may be related to the wide scope of the test.

For the validity of the test, opinions of experts in the field of measurement and evaluation and language are asked. For content validity, 14 alternative measurement and evaluation concepts in the 2017 curriculum are included in the test in line with expert opinions. For criterion validity, the correlation of the two-tier test (ODKT) developed by Arık (2006) and AMEMIT has been examined. There is a statistically insignificant relationship between the first tiers in the tests. r=0.25, *p*>.05. The reason for the insignificant relationship can be that the first tier of AMEMIT consisted of 4 options and the first tier of ODKT consisted of 2 options. There is a moderate, positive and significant relationship between the tiers in which the justifications are included in the tests. r=0.53, *p*<.05. The reason for the increase in the relationship at this stage can be that the third tier of AMEMIT and ODKT consisted of 4 options. Additionally, the lack of a high correlation between the tests can be due to the content of ODKT being related to measurement and evaluation, whereas AMEMIT is related to alternative measurement and evaluation. Criterion validity has not been provided for the first tier but has been provided in the third tier. For the construct validity of the test, the relationship between the teachers' correct answer scores and their reliability scores has been examined. Explanations made by the respondents about how they answered the questions and what they thought during or after answering gives important knowledge in explaining the structure of the test (Baykul, 2015). First, the Pearson correlation coefficient is calculated to determine the relationship between the first and third tiers of AMEMIT. A moderate, positive and significant relationship was found between these tiers. r=0.58, *p*<.05. Secondly, biserial correlation is calculated to determine the relationship between the first and second tiers of AMEMIT. These correlation values vary between 0.20 and 0.62. According to these values, there is a low positive correlation for just one item and a moderate positive correlation for all other items. Accordingly, those who are sure in the reliability tier of the test have gotten higher scores from the test. Also for construct validity, the rates of false positive (correct with false reason) and false negative (false with correct reason) have been examined. The false positive rate is 9.4% and the false negative rate is 13.3%. In these tests, the false negative and false positive rate should be less than 10% (Hestenes and Holloun, 1995).

Coding and Analysis of Data

The answers given to AMEMIT have been analyzed with the help of MS Excel program. In the coding made in Excel, 1 is used for the correct answer and 0 for the wrong answer in the first and third tiers. In the reliability level in the second and fourth tiers, 1 is used for sure and 0 for not sure. There are 16 different combinations according to this coding system. With the help of these combinations, teachers' levels of scientific knowledge, misconceptions, false positive, false negative and lack of knowledge are identified. For this, frequency and percentage statistics have been used.

Result / Findings

Findings Regarding Teachers' Misconceptions in the Field of Alternative Measurement and Evaluation

The frequencies and percentages of Scientific Knowledge (SN), Misconception (M), False Positive (FP), False Negative (FN) and Lack of Knowledge (LK) calculated in line with the answers given by the teachers to item 1, which was given as an example during the test development process, are given in Table 2.

Table 2. Frequency and Percentages Related to the Item

Response Pattern	Explanation	Frequency	Percentage
0000	LK12	14	3.9%
0001	LK11	6	1.7%
0010	LK9	6	1.7%
0011	LK8	6	1.7%
0100	LK10	11	3.1%
0101	M	74	20.6%
0110	LK7	4	1.1%
0111	FN	62	17.2%
1000	LK6	13	3.6%
1001	LK5	9	2,5%
1010	LK3	5	1.4%
1011	LK2	4	1.1%
1100	LK4	6	1.7%
1101	FP	64	17.8%
1110	LK1	6	1.7%
1111	SN	70	19.4%
Total		360	100.0%
Total LK		90	25.0%

SN=Scientific Knowledge, M=Misconception, FP= False Positive, FN= False Negative and LK=Lack of Knowledge

As seen in Table 2, 25.0% of the teachers have the most lack of knowledge in the concept of portfolio. In the combination of lack of knowledge, it has the most LK12 and this rate constitutes 3.9% of the whole group. This finding shows that although teachers do not know the concept of portfolio and for what purpose the portfolio is used, they are not sure in the reliability tier either. Afterwards, secondly, they have misconceptions with

20.6%. 19.4% of teachers know the concept of portfolio and for what purpose it is used, 17.8% have false positive and 17.2% have false negative the findings of the other items in the test are given in Table 3.

Table 3. Findings Regarding the Items in the Test

Item	(SN)		(M)		(LK)		(FP)		(FN)	
	f	%	f	%	f	%	f	%	f	%
1	70	19.4	74	20.6	90	25.0	64	17.8	62	17.2
2	31	8.6	75	20.8	122	33.9	22	6.1	110	30.6
3	21	5.8	218	60.6	99	27.5	13	3.6	9	2.5
4	56	15.6	36	10.0	185	51.4	72	20.0	11	3.1
5	120	33.3	68	18.9	102	28.3	35	9.7	35	9.7
6	92	25.6	51	14.2	116	32.2	45	12.5	56	15.6
7	76	21.1	49	13.6	87	24.2	11	3.1	137	38.1
8	29	8.1	76	21.1	198	55.0	35	9.7	22	6.1
9	122	33.9	71	19.7	100	27.8	30	8.3	37	10.3
10	26	7.2	60	16.7	177	49.2	41	11.4	56	15.6
11	65	18.1	82	22.8	136	37.8	37	10.3	40	11.1
12	73	2.3	58	16.1	159	44.2	24	6.7	46	12.8
13	51	14.2	50	13.9	184	51.1	52	14.4	23	6.4
14	146	40.6	49	13.6	132	36.7	15	4.2	18	5.0
15	42	11.7	68	18.9	176	48.9	16	4.4	58	16.1
Avarage	18.9		20.1		38.2		9.4		13.3	

When Table 3 is examined, teachers generally have misconceptions in every concept. For some concepts, their scientific knowledge and lack of knowledge are found at a higher level. Scientific knowledge predominates in the concepts of feedback (item 9) and peer assessment (item 14), which teachers use in their classes and are more familiar with. Lack of knowledge comes to the fore in concepts such as structural grid (item 8), word association test (item 10) and branching tree (item 13), which they do not use in their classes. When the average values of the test have been examined, it is concluded that the teachers mostly have misconceptions after the lack of knowledge in the field of alternative measurement and evaluation.

Discussion and Conclusion

In this research, it is aimed to develop a four-tier test to identify teachers' misconceptions in the field of alternative assessment and evaluation and to identify their misconceptions. The reliability and validity studies of the test have been carried out and it is concluded that the test is a reliable and valid measurement tool that identifies the misconceptions of teachers in the field of alternative assessment and evaluation. For each item, more than 10% lack of knowledge and misconceptions are found.

The false negative rate in the test is 13.3%. In these tests, the false negative rate should be less than 10%. Minimizing false positive and false negative rates is a big problem. In addition, false negative of more than 10% for some items may also be caused by the lack of attention of the respondent (Hestenes and Holloun, 1995).

The lack of knowledge is found in the most structural grid concept with 55.0%. This finding shows that the majority of teachers do not know what the concept of structural grid is and for what purpose it is used. In their research, Çermik (2011) and Karamustafaoğlu et al., (2012) concluded that the concept of structural grid is the least known and least used method by teachers. At the same time, it is concluded in some studies that teachers never used the structural grid method (Karalok, 2014; Özenç, 2013).

The misconception is seen mostly in the concept of performance task, the 3rd item, with 60.6%. In some studies, it has been concluded that the majority of teachers frequently use the concept of performance task and they see themselves as competent in this field (Acar and Anıl, 2009; Aksu, 2013; Çermik, 2011; Duran et al., 2013; Okur, 2008; Özdemir, 2010; Özenç, 2013). The fact that teachers consider themselves competent in this method and frequently use this method in their classes does not mean that they have sufficient knowledge in this method or that they do not have misconceptions. In some studies (Çalışkan, 2009; Gelbal and Kelecioğlu, 2007; Orhan, 2007), it has been concluded that although teachers mostly prefer traditional measurement and evaluation methods such as paper-pencil test to make definite judgments for students, the knowledge level of teachers in performance task method is not at the desired level (Acar and Anıl, 2009). The widespread use of traditional measurement and evaluation methods in education may cause teachers to have misconceptions.

Scientific knowledge is mostly seen in the peer assessment, the 14th item, with 40.6%. This finding shows that the majority of teachers know the concept of peer assessment and what it is used for. İn their study of primary school teachers' proficiency level for alternative assessment methods, concluded that 57.5% of the teachers felt competent in the concept of peer assessment and they used peer assessment very rarely with 32.5%. In his study determining the traditional and alternative assessment methods used by regular classroom teachers in the classroom, Özenç (2013) stated that 4 out of 9 teachers who were observed had used the peer assessment method.

False negative is mostly seen in the group evaluation concept, the 7th item, with 38.1%. Teachers have a false negative by calling the concept of group assessment as peer assessment and specifying the purpose for which peer assessment is used. False positive is mostly seen in the observation, the 4th item, with 20%. Most of the teachers have false negative because of saying that the student will take an active role in the process while observing.

Since the four-tier tests have a reliability level and 16 different answer combinations for each tier, they are more effective than other tests in distinguishing between misconception, scientific knowledge, lack of knowledge, false positive and false negative. Thus, it may be advantageous to use a four-tier test in different misconception studies. Although alternative assessment and evaluation methods have been in the curriculum for many years, teachers have lack of knowledge and misconceptions in these methods. Trainings for teaching these methods can be organized and transferred to classroom teaching practices. More class hours can be allocated to alternative measurement and evaluation methods in the measurement and evaluation course for teacher candidates. Additionally, in order to pass the course successfully, methods such as portfolio, self-assessment, peer assessment can be used in the process. In this way, teacher candidates can learn the methods through practical application.

References

Acar, M. and Anıl, D. (2009). Sınıf öğretmenlerinin performans değerlendirme sürecindeki değerlendirme yöntemlerini kullanabilme yeterlikleri, karşılaştıkları sorunlar ve çözüm önerileri. *TÜBAV Bilim Dergisi, 2*(3), 354-363. https://dergipark.org.tr/tr/download/article-file/799616_

Akbaş, U., Gürkan, B., and Büyüköztürk, Ş. (2018). Ortaokul matematik öğretim programlarının ölçme değerlendirme yaklaşımları. In M. F. Özmantar, H. Akkoç, B. Kuşdemir Kayıran, M. Özyurt (Ed.), *Ortaokul matematik öğretim programları tarihsel bir inceleme* (349-365). Ankara: Pegem Akademi.

Aksu, Ö. (2013). *Biyoloji öğretmenlerinin uyguladıkları alternatif ölçme ve değerlendirme tekniklerinin değerlendirilmesi ve öğretmen-öğrenci görüşleri*. Unpublished doctoral thesis, Gazi University Institute of Education Sciences.

Arık, S. R. (2006). *İlköğretim öğretmenlerinin ölçme ve değerlendirme alanındaki kavram yanılgılarının belirlenmesi*. Unpublished master’s thesis, Ankara University Institute of Education Sciences.

Bahar, M., Nartgün, Z., Durmuş, S., and Bıçak, B. (2015). *Geleneksel-tamamlayıcı ölçme ve değerlendirme teknikleri öğretmen el kitabı* (7th ed.). Ankara: Pegem Akademi.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması* (3rd ed.). Ankara: Pegem Akademi.

Bekiroğlu, F. O. (2004). *Ne kadar başarılı? Klasik ve alternatif ölçme değerlendirme yöntemleri: Fizikte uygulamalar*. Ankara: Nobel Yayıncılık.

Burgoon, J. N., Heddle, M. L., and Duran, E. (2011). Re-examining the similarities between teacher and student conceptions about physical science. *Journal of Science Teacher Education, 2*(22), 101-114. https://sci-hub.se/10.2307/43156591

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., and Demirel, F. (2018). *Bilimsel araştırma yöntemleri* (24th ed.). Ankara: Pegem Akademi.

Caleon, I., and Subramaniam, R. (2010a). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education, 32*(7), 939-961. Doi: 10.1080/09500690902890130

Caleon, I., and Subramaniam, R. (2010b). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education, 40*(3), 313-337. Doi: 10.1007/s11165-009-9122-4

Çalışkan, İ. (2009). *Fen ve teknoloji öğretmen adaylarının tamamlayıcı ölçme ve değerlendirme yaklaşımlarını kullanma becerileri ile fen ve teknoloji öğretmen ve öğretmen adaylarının bu yaklaşımlarla ilgili görüşleri hakkında durum belirleme.* Unpublished doctoral dissertation, Hacettepe University , Institute of Education Sciences.

Çermik, F. (2011). *Yeni ilköğretim programlarının öngördüğü tamamlayıcı ölçme değerlendirme teknikleri hakkındaki öğretmen görüşlerinin değerlendirilmesi*. Unpublished master's thesis, Fırat University Institute of Education Sciences.

Demirbilek, S. (2015). *Öğretmen adaylarının eğitimde ölçme ve değerlendirme dersindeki kavram yanılgılarının incelenmesi*. Unpublished master's thesis, Hacettepe University Institute of Education Sciences.

Duran, M., Mıhladız, G., and Ballıel, G. (2013). İlköğretim öğretmenlerinin alternatif değerlendirme yöntemlerine yönelik yeterlik düzeyleri. *Mehmet Akif Ersoy Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi*, *2*(2), 26-37. Retrieved from https://dergipark.org.tr/tr/download/article-file/207766

Eryılmaz, A., and Sürmeli, E. (2002). *Üç aşamalı sorularla öğrencilerin ısı ve sıcaklık konularındaki kavram yanılgılarının ölçülmesi*. https://users.metu.edu.tr/eryilmaz/TamUcBaglant.pdf

Fratiwi, N. J., Kaniawati, I., Suhendi, E., Suyana, I. and Samsudin, A. (2017). The transformation of two-tier test into fourtier test on Newton's laws concepts. *AIP Conference Proceedings, 18*(8), 1848. Doi: 10.1063/1.4983967

Gelbal, S., and Kelecioğlu, H. (2007). Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkında yeterlik algıları ve karşılaştıkları sorunlar. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 33, 135-145. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1017-published.pdf

Görkemli-Taban, T. (2017). *Fen bilgisi öğretmen adaylarının sıvı basıncı konusundaki kavram yanılgılarının dört aşamalı tanı testi ile belirlenmesi*.

Unpublished master's thesis, Necmettin Erbakan University Institute of Education Sciences.

Güneş, B. (2005). Bilimsel hatalar ve kavram yanılgıları. In R. Yağbasan, (Ed.), Konu alanı ders kitabı inceleme kılavuzu fizik (59-114). Ankara: Gazi Kitapevi.

Hasan, S., Bagayoko, D., and Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education, 34*(5), 294-299. Doi: 10.1088/0031-9120/34/5/304

Hestenes, D., and Halloun, I. (1995). Interpreting the force concept inventory: A response to Huffman and Heller. *The Physics Teacher, 33*(8), 502-506. Doi: 10.1119/1.2344278

Kaltakçı, D. (2012). *Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics*. Unpublished doctoral thesis, Middle East Technical University Social Sciences Institute.

Kaltakçı-Gürel, D., Eryılmaz, E., and McDermott, L. C. (2015). A rewiev and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(5), 989-1008. Doi: 10.12973/eurasia.2015.1369a

Kanlı, U. (2015). Using a two-tier test to analyse students' and teachers' alternative concepts in astronomy. *Science Education International, 26*(2), 148-165. https://files.eric.ed.gov/fulltext/EJ1064041.pdf

Karadeniz-Bayrak, B. (2013). Using two-tier test to identify primary students' conceptual understanding and alternative conceptions in acid base. *Mevlana International Journal of Education, 3*(2), 19-26. Doi: 10.13054/mije.13.21.3.2

Karalök, S. (2014). *Ortaokul matematik öğretmenlerinin matematik dersi tamamlayıcı ölçme değerlendirme tekniklerine ilişkin profilleri*. Unpublished master's thesis, Pamukkale University, Institute of Education Sciences.

Karamustafaoğlu, S., Çağlak, A., and Meşeci, B. (2012). Alternatif ölçme değerlendirme araçlarına ilişkin sınıf öğretmenlerinin öz yeterlilikleri. *Amasya Üniversitesi Eğitim Fakültesi Dergisi, 1*(2), 167-179. https://dergipark.org.tr/tr/download/article-file/19598

Kılınç, S. (2017). *Fen bilgisi öğretmen adaylarının yoğunluk konusundaki kavram yanılgılarının dört aşamalı tanı testi ile belirlenmesi*. Unpublished master's thesis, Necmettin Erbakan University Institute of Education Sciences.

Kutlu, Ö., Doğan, C. D., and Karakaya, İ. (2017). *Ölçme ve değerlendirme performansa ve portfolyoya bağlı durum belirleme* (5th ed.). Ankara: Pegem Akademi.

Leonard, M., Kalinowski, S. T. and Andrews, T. C. (2014). Misconceptions yesterday, today, and tomorrow. *CBE Life Sciences Education, 13*(2), 179-186. Doi: 10.1187/cbe.13-12-0244

Lin, Y. C., Yang, D. C. and Li, M. N. (2015). Diagnosing students' misconceptions in number sense via a web-based two-tier test. *Eurasia Journal of Mathematics, Science & Technology Education, 12*(1), 41-55. Doi: 10.12973/eurasia.2016.1420a

MEB. (2017). *İlköğretim Türkçe dersi öğretim programı ve kılavuzu, 1-8. sınıflar.* Ankara: MEB. https://web.deu.edu.tr/ilyas/ftp/turkce2017.pdf

MEB. (2005). *İlköğretim okulu ders programları ve öğretim kılavuzları, 1-5. sınıflar.* Erzurum: Yakutiye Yayıncılık.

MEB. (2009). *Millî Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı ilköğretim matematik dersi 6-8. sınıflar öğretim programı ve kılavuzu*. https://akademik.adu.edu.tr/ad/egitim/mat/webfolders/Mat_6-8_2009.pdf

Meşin, M. Z. (2019). *Fen bilgisi öğretmen adaylarının gaz kanunları ile ilgili kavram yanılgılarının dört aşamalı test ile belirlenmesi*. Unpublished master's thesis, Necmettin Erbakan University, Institute of Education Sciences.

Milenković, D. D., Hrin, T. N., Segedinac, M. D. and Horvat, S. (2016). Development of a three-tier test as a valid diagnostic tool for identification of misconceptions related to carbohydrates. *Journal of Chemical Education, 93*(9), 1514-1520. Doi: 10.1021/acs.jchemed.6b00261

Okur, M. (2008). *4. ve 5. sınıf öğretmenlerinin fen ve teknoloji dersinde kullanılan alternatif ölçme ve değerlendirme tekniklerine ilişkin görüşlerinin belirlenmesi.* Unpublished master's thesis, Zonguldak Karaelmas University, Social Sciences Institute.

Orhan, A. T. (2007). *Fen eğitiminde alternatif ölçme ve değerlendirme yöntemlerinin ilköğretim öğretmen adayı, öğretmen ve öğrenci boyutu dikkate alınarak incelenmesi*. Unpublished doctoral thesis, Gazi University, Institute of Education Sciences.

Önsal, G. (2016). *Özel görelilik kuramıyla ilgili kavram yanılgılarını belirlemeye yönelik dört aşamalı bir testin geliştirilmesi ve uygulanması.* Unpublished master's thesis, Gazi University, Institute of Education Sciences.

Özdemir, S. M. (2010). İlköğretim öğretmenlerinin alternatif ölçme ve değerlendirme araçlarına ilişkin yeterlikleri ve hizmet içi eğitim ihtiyaçları. *Türk Eğitim Bilimleri Dergisi, 8*(4), 787-816. https://dergipark.org.tr/tr/download/article-file/256227

Özenç, M. (2013). *Sınıf öğretmenlerinin alternatif ölçme ve değerlendirme yeterliklerinin incelenmesi.* Unpublished doctoral thesis, Marmara University Institute of Education Sciences.

Özmantar, M. F., Ağaç, G., Yılmaz, G. and Özbey, N. (2018). Cumhuriyet dönemi ortaokul matematik öğretim programlarına genel bir bakış. In M. F. Özmantar, H. Akkoç, B. Kuşdemir Kayıran, M. Özyurt (Ed.), *Ortaokul matematik öğretim programları tarihsel bir inceleme* (29-75). Ankara: Pegem Akademi.

Peşman, H. and Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research, 103*(3), 208-222. Doi: 10.1080/00220670903383002

Salvucci, S., Walter, E., Conley, V., Fink, S., and Saba, M. (1997). *Measurement error studies at the national center for education statistics (NCES).* Washington: D. C. U. S. Department of Education Publishers.

Sheppard, K. (2006). High school students' understanding of titrations and related acid-base phenomena. *Chemistry Education Research and Practice, 7*(1), 32-45. Doi: 10.1039/b5rp90014j

Smith, J. P., DiSessa, A. A., and Rochelle, J. (1994). Misconceptions reconceived: a constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.
Doi: 10.1207/s15327809jls0302_1

Sreenivasulu, B., and Subramaniam, R. (2013). University students' understanding of chemical thermodynamics. *International Journal of Science Education, 35*(4), 601-635.
Doi: 10.1080/09500693.2012.683460

Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple-choice items. *Research in Science Education, 16*(1), 199-207.
Doi: 10.1007/BF02356835

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159-169.
Doi: 10.1080/0950069880100204

Ün-Açıkgöz, K. (2003). *Aktif öğrenme* (3rd ed.). İzmir: Eğitim Dünyası Yayınları.

Üztemur, S. S. (2013). *Sosyal bilgiler öğretmenlerinin ölçme ve değerlendirme alanındaki kavram yanılgıları ve öz-yeterlik inançlarının incelenmesi*. Unpublished master's thesis, Celal Bayar University, Social Sciences Institute.

Yang, D. C. (2019). Development of a three-tier number sense test for fifth-grade students. *Educational Studies in Mathematics, 101,* 405-424.
Doi: 10.1007/s10649-018-9874-8.