⌈Araştırma Makalesi / Research Article⌋

Kastamonu Eğitim Dergisi
Kastamonu Education Journal
Ocak 2020 Cilt:28 Sayı:1
kefdergi.kastamonu.edu.tr

Başvuru Tarihi/Received: 11.10.2019
Kabul Tarihi/Accepted: 09.01.2020
DOI: 10.24106/kefdergi.3980

Enformatik Dersi için Başarı Testi Geliştirme Çalışması: Güvenilirlik ve Geçerlilik İşlemleri[1]

Achievement Test Development Study for Informatics Lesson: Reliability and Validity

Murat MERİÇELLİ[2] & Tolga GÜYER[3]

Öz

Bu çalışmanın amacı geçerlilik ve güvenilirlik çalışmalarının yapıldığı bir enformatik dersi başarı testi geliştirmektir. İlk olarak, 25 alt konuya, her birine 2 katı soru olacak şekilde, bir içerik ağacında 80 çoktan seçmeli soru belirlenmiştir. 8 uzmana sorular sorulmuş ve sorular ve soruların ait olduğu birimler üzerinde kapsam geçerlilik indeksi hesaplanmıştır. Test için hesaplanan kapsam geçerlilik endeksi 0,91'dir. Başarı testi 21 BÖTE 3. sınıf öğrencisine uygulanmıştır. TAP programı yardımıyla madde analizi yapılmıştır. Kuder-Richardson-20 Güvenilirlik katsayısı değeri 0,858 olarak hesaplanmıştır. Her alt konuda madde ayırt edicilik ve madde zorluk değerleri uygun olan maddeler belirlenerek toplamda 40 soruluk teste ulaşılmıştır. Elde edilen testte ortalama madde zorluk değeri 0,638 olarak bulunmuştur. Bu durum madde zorluğunun orta düzeyde olduğu anlamına gelmektedir. Ortalama madde ayırt ediciliği 0,444 olarak bulunmuştur. Bu değer madde ayırt edicilik özelliğinin çok iyi düzeyde olduğu anlamına gelmektedir. 40 çoktan seçmeli soruya indirgenmiş geçerlilik ve güvenilirlik süreçleri yapıldığı için, test güvenilir ve geçerlidir, farklı enformatik dersleri için bir başarı testi olarak kullanılabilir.

Anahtar Kelimeler: başarı testi, içerik geçerlilik indeksi, güvenilirlik, enformatik dersi

Abstract

The purpose of this study is to develop an Informatics course achievement test that validity and reliability studies have been carried out. First, 80 multiple choice questions were determined in a content tree, which is 2 times for each sub-subject(n=25). Questions were asked to 8 experts and the content validity index was calculated on the questions and the units to which the questions belong. The content validity index calculated for the test was 0.91. The achievement test was applied to 21 CEIT 3rd-grade students. Item analysis was carried out with the help of the TAP program. Kuder-Richardson-20 Reliability coefficient value was calculated as 0.858. In each sub-item, item discrimination and item difficulty values were determined and a total of 40 questions were reached as a result. In the obtained test, the mean item difficulty value was found to be 0.638. This means that the item difficulty is average. The mean item discrimination was found to be 0.444. This value means that the item discrimi-nation is very good. Since the validity and reliability pro-cesses reduced to 40 multiple-choice questions have been conducted, the test is reliable and valid can be used as an achievement test for different informatics courses.

Keywords: achievement test, content validity index, reliability, informatic lesson

1 This study has been produced under the first author's Ph.D thesis at Gazi University.

2 Kastamonu Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, Kastamonu, Türkiye, https://orcid.org/0000-0003-0168-3221
3 Gazi Üniversitesi, Gazi Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, Kastamonu, Türkiye, https://orcid.org/0000-0001-9175-5043

Atıf / Citation: Meriçelli, M., & Güyer, T. (2020). Achievement Test Development Study for Informatics Lesson: Reliability and Validity. *Kastamonu Education Journal, 28*(1), xxx-xxx. doi:10.24106/kefdergi.3980

1. Introduction

Nowadays, wherever human beings exist at home, at work, and school, Information & Communication Technology (ICT) also finds its application area and has become a part of daily life. It is inevitable that teaching is shaped according to 21st century standards and is widely used in teaching activities. In this regard, the effects of different ICTs in teaching and learning have been researched and their effects have been found in many experi-mental studies (Toro & Joshi, , 2012; McDougall & Jones, 2006; Eng, 2005). For this reason, we observe that the use of ICT in learning environments is increasing by policymakers and researchers with the help of different projects (e.g: "FATİH Project" in Turkey). Thus, it is known that candidate teachers should undergo appropriate ICT teaching. The necessity, general framework of basic computer education in higher education institutions and the necessity of these courses to be given by Informatics Departments were determined by Higher Education Council (YÖK) and communicated to universities for execution. It is clearly stated that basic computer courses should be carried out with two compulsory courses and if necessary, it is stated that universities can supplement these compulsory courses with elective courses. It can be understood that in the evaluation of informatics cours-es given at different universities, a valid and reliable achievement test will be needed.

Considering the scope of ICT for this teaching, there will be a large sub-application pool (Toro & Joshi, 2012 ; Kozma, 2005). Subjects like word processor, spreadsheet, presentation etc. could not be enough for an ICT lit-erate teacher and the focal points of the application areas can be changed (Oliver, 2002). Therefore, a content tree was created for Information Technologies II course considering the needs of teacher candidates. In addition to that, there was a need to measure the level of learning of ICT-related contents at an adequate level. This meas-urement could be done with exams/ achievement tests.

Exams are important measurement tools that can take place at every stage of teaching. Measurement is the ex-pression of a feature by number symbols or adjectives (Sönmez & Alacapınar, 2016). The evaluation allows the decision to determine the quality of the measurement results measured according to a criterion (Sönmez & Alacapınar, 2016; Yıldırım, 1999). It will be possible to evaluate the measurements obtained from exams. This means that it is necessary to decide whether learning is enough (Yaşar, 2017). The examination in which this decision will be made must meet certain criteria. In order to measure and evaluate student achievement, true-false tests, matching test, fill in the blank tests, short-answer tests, open-ended questions and multiple-choice tests any other form of the test are used in different cases. Each test form has its pros and cons according to different con-ditions. Multiple-choice tests come into prominence in terms of their ability to be applied to large masses, easy to carry out measurement and evaluation and to correct misconceptions (Çakan, 2017; Treagust, 1986). The devel-opment of this achievement test was also chosen for one of the most widely used multiple choice tests for the following reasons:

- Evaluating is the most objective exam type.
- The answering and scoring time is advantageous because it is short.
- Because it consists of many questions, it is a comprehensive, valid and reliable exam.
- Many statistical procedures can be performed with the data obtained in this type of exam.
- Easily applicable to large groups
- Can be used in all teaching levels

The instruments of measurement must provide two qualities: validity and reliability. The reliability is that a measuring instrument measures the same result every time (Carmines & Zeller, 1979; Crocker & Algina, 1986; Şencan, 2005). Validity is related to whether the measuring tool measures what it wants to measure (Büyüköztürk, Akgün, Demirel, Karadeniz, & Çakmak, 2015; Garrett, 1937). A measurement must be reliable to be valid. However, although reliability is a necessary condition for validity, it is not an adequate condition. High validity can also mean high reliability, but not vice versa. In other words, high reliability gives no information about validity. Also, sometimes the steps to make the test reliable may conflict with the steps to make the test valid (Karasar, 2005).

In the process of developing multiple-choice tests, steps such as reliability item analysis and content validity were run in different achievement test development studies. On the other hand, it can be said that the procedures performed to ensure validity should be objective too (Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003). Because of that, in this study also validity studies were done based on quantitative approach.

This study aims to develop a valid and reliable measurement tool for the Informatics II course. In addition to the

reliability, the 80-item achievement test aims to provide validity based on the measurement and the content, and methods are discussed in accordance with this purpose.

2. Method

In this study, 40 multiple choice questions with 5 options for 8-week course content were evaluated to provide validity and reliability. For the 8-week course content, it is planned to have an equal number of questions to meet the content validity. In the pilot implementation, 2 questions were written for each sub-content.

Table 1. Test Items for Achievement Test based on Content Tree

Week	Subject	Sub Subject	Count	Item Number
1	Excel I	What is Excel?	2	5, 7
		Excel Layout	2	1, 4
		Excel Home Tab	4	6, 8, 9, 10
2	Excel II	Excel File Tab	2	2, 3
		Data Tab	6	11, 12, 16, 17, 19, 20
		Insert Tab	4	13, 14, 15, 18
3	Excel III	Cells & Cell Range	2	22, 23
		Arithmetic Formulas	2	21, 27
		Arithmetic Operators	2	25, 26
		Conditional Formulas	4	24, 29, 30, 28
4	Computers & OS	How is it Works?	4	32, 33, 34, 35
		History	4	31, 36, 39, 40
		Operating Systems	2	37, 38
5	Internet & Web	Internet	2	48, 49
		Web Developing Process	2	47, 50
		Web Versions	2	41, 42
		Web 2.0 Tools	4	43, 44, 45, 46
6	Cloud Systems	Cloud Technology	4	51, 59, 52, 58
		Cloud Technology Tools	6	53, 54, 55, 56, 57, 60
7	CBL	Computer Based Learning	4	64, 65, 66, 68
		Instructional Software	2	67, 70
		Types of Instructional Software	4	61, 62, 63, 69
8	Distance Education	Distance Education	2	71, 80
		Advantages & Disadvantages	4	72, 73, 75, 76
		Tools	4	74, 77, 78, 79

80 questions were obtained as a result of creating double questions for each sub-content. Before practice, these questions need to be determined to provide validity. For this purpose, the content validity index, which was developed by Lawshe (1975) to provide content validity and which gives more systematic results for determining the content validity, was preferred.

An expert group consisting of 8 experts, 2 of whom are Information Technology teachers and 6 of whom are academicians in Computer Education and Instructional Technologies Education have been utilized to operate this method. The content tree and the questions were presented to the expert group in written form. For each question, they were asked to select one of the options “item measures the targeted structure”, “item related to the structure but unnecessary” and “item does not measure the targeted structure”, and explain if there are sections that they consider it necessary to be corrected. The response as “Item measures the targeted structure” means the experts who approve test items could be used in Formula 1.

$$CVR = \frac{NA}{N/2} - 1 \qquad formula(1)$$

NA: Number of experts that approve test items could be used

N: The total count of experts who had contributed

CVR: Content Validity Rates

The CVR values in formula 1 were calculated for each question. While the overall arithmetic mean of the CVR values gives the CVI value for the test, the arithmetic mean for the units gives the CVI value for the units (Yurdugül, 2005). The CVI is calculated based on the response of each expert to the question. The change in the number of experts also changes the minimum targeted value for the CVI. Table 2 shows the expected minimum CVI values for dif-ferent numbers of experts.

Table 2. Minimum values of CVI based on expert numbers (Veneziano & Hooper, 1997)

Number of Experts	Min Value	Number of Experts	Min Value	Number of Experts	Min Value
5	0.99	10	0.62	15	0.49
6	0.99	11	0.59	16	0.42
7	0.99	12	0.56	17	0.37
8	0.78	13	0.54	18	0.33
9	0.75	14	0.51	19	0.31

21 students of the Department of Computer Education and Instructional Technology (CEIT) who attended the Special Teaching Methods II course in the 2018-2019 spring semester participated in this study. The reason to select CEIT students is that they were currently the only department with enough knowledge of the content. Informatics is taught in many cohorts; however, due to drastic changes in the curriculum, students in other departments do not have enough knowledge of the content.

According to the classical test theory, there are 3 basic statistics should be paid attention to the test. These are item difficulty, item distinctiveness index and reliability coefficient (Baykul, 2015; Crocker & Algina, 1986; Verhelst, 2014). There are different assessment ranges for different substances according to different experts. The use of the evaluation ranges shown in Table 3 was preferred in the study.

Table 3. Difficulty Levels of Items (Sözbilir, 2010)

Difficulty of Item(p)	Assessment of item
0.00 - 0.19	Very difficult
0.20 – 0.34	Difficult
0.35 - 0.65	Average
0.65 - 0.79	Easy
0.80- 1.00	Very easy

The ranges of values to be used in the evaluation of the item discrimination are shown in Table 4.

Table 4. Distinctiveness Criteria of Items (Özçelik, 1992)

Distinctiveness (r)	Assessment of item
0.19 and lower	Unacceptable
0.20 – 0.29	Must be revised
0.30 - 0.39	Good Acceptable
0.40 and higher	Very Good Acceptable

For reliability, test-retest, parallel forms, two semi-tests KR-20, and Cronbach Alpha can be used. It is preferred to use KR-20 to determine the reliability of the item analysis (Büyüköztürk vd., 2015). The equation used in the calculation of KR-20 is shown in Formula 2.

$$KR_{20} = \frac{K}{K-1}\left[1 - \frac{\sum pq}{S_x^2}\right] \qquad formula(2)$$

K = Number of test items

p =Item difficulty

q = 1-p

S^2_x = Variance of the test

Data analysis of substances was carried out by using Test Analysis Program (TAP) (Brooks & Johanson, 2003). The TAP allows test statistics more easily and quickly (Ayhan, 2010). Item difficulty index and item discrimination index values in the data analysis were performed. Furthermore, Kuder-Richardson 20 value, average, variance, standard deviation was obtained.

3. Findings

3.1 Validity of the test

Table 5. Content Validity Ratios and Content Validity Indexes

Week/Unit	CVI	Week/Unit	CVI
1	1.00	5	0.95
2	0.90	6	0.875
3	0.95	7	0.852
4	0.975	8	0.80
Content Validity Index Criteria for 8 Experts			0.78*
Content Validity Index (CVI)			0.91

The required level of content validity index is 0.78 for 8 experts. It is seen in Table 5 that 0.78 (Veneziano & Hooper, 1997), which is the benchmark value for 8 experts, is provided according to the calculated validity index (CVI = 0.91). Also, the content validity index calculated for each unit is above the criterion value. Besides, in the light of the feedback from the experts, arrangements were made on the 21 question roots and options.

3.2 Normality of the test

Shapiro-Wilks was used to determine the normal distribution. If the p-value is greater than 0.05, the hypothesis is accepted, and the distribution does not differ significantly from the normal distribution. According to the results of the Shapiro-Wilks test, it was concluded that the data were distributed normally and there was no significant difference between the normal distribution. (p = 0.587 s = 0.963)

3.3 Test Analysis of achievement test

The average score for the test was 46.429 variance 92.626, standard deviation of 9.624. Kuder-Richardson-20 Reliability coefficient value was calculated as 0.858. Accordingly, KR-20 value indicates that enough reliability of the test score is above 0.70.

Table 6. Descriptive Statistics

Number of Items	Mean	Variance	Std. Dev	KR-20
80	46.429	92.626	9.624	0.858

Besides displaying the # sign through the TAP in question (p <0.2 or p> 0.95, D <0, pbis <0, adjpbis <0) have been identified as potential problems. It was determined that there are potential problems in 25 questions overall. 10 of them are below 0.20, 5 of them are below 0, 5 of them are below 0, and 10 of them are below 0.

Table 7. Item Analysis Result of Achievement Test

Item No	pj	rjx	Upper Group Correct Answer Score	Lower Group Correct Answer Score	Difficulty	Discrimination
1	0.43	0.5	4	1	average	very good
2	0.71	0.5	5	2	easy	very good
3	0.95	0.17	6	5	very easy	very low
4	0.95	0.17	6	5	very easy	very low
5	0.95	0	6	6	very easy	very low
6	0.9	0.33	6	4	very easy	good
7	0.81	0.67	6	2	very easy	very good
8	0.86	0.33	6	4	very easy	good
9	0.48	0.5	4	1	average	very good
10	0.38	0.5	4	1	average	very good
11	0.67	0.33	5	3	easy	good
12	0	0	0	0	very difficult	very low
13	0.24	0.17	3	2	difficult	very low
14	0.81	0.33	6	4	very easy	good
15	0.24	-0.17	2	3	difficult	very low
16	0.71	0.5	6	3	easy	very good
17	0.67	0.33	5	3	easy	good
18	0.71	0.67	6	2	easy	very good
19	0.19	0.33	2	0	very difficult	good
20	0.14	0.33	2	0	very difficult	good
21	0.43	0.33	4	2	average	good
22	0.81	0.5	6	3	very easy	very good
23	0.67	-0.17	3	4	easy	very low
24	0.76	0.17	5	4	easy	very low
25	0.62	0.17	4	3	average	very low
26	0.81	0.33	6	4	very easy	good
27	0.05	0	0	0	very difficult	very low
28	0.71	0.83	6	1	easy	very good
29	0.86	0.33	6	4	very easy	good
30	0.62	0.33	6	4	average	good
31	0.1	0.17	1	0	very difficult	very low
32	0.9	0	6	6	very easy	very low
33	0.9	0.33	6	4	very easy	good
34	0.9	0.33	6	4	very easy	good
35	1	0	6	6	very easy	very low
36	0.14	-0.33	0	2	very difficult	very low
37	0.19	0.33	2	0	very difficult	good
38	0.24	0.5	3	0	difficult	very good
39	0.57	-0.17	4	5	average	very low
40	0.62	0.5	5	2	average	very good
41	0.76	0.17	5	4	easy	very low
42	0.71	0.33	6	4	easy	good
43	0.81	0	5	5	very easy	very low
44	0.52	0.33	4	2	average	good
45	0.71	0	5	5	easy	very low
46	0.38	0.33	3	1	average	good

Item No	pj	rjx	Upper Group Correct Answer Score	Lower Group Correct Answer Score	Difficulty	Discrimination
47	0.81	0.5	6	3	very easy	very good
48	0.29	0.17	2	1	difficult	very low
49	0.48	0.5	4	1	average	very good
50	0.24	0.33	3	1	difficult	good
51	0.43	0.33	4	2	average	good
52	0.67	0	5	5	easy	very low
53	0.48	0.5	5	2	average	very good
54	0.57	0.67	5	1	average	very good
55	0.43	0.33	3	1	average	good
56	0.19	0	1	1	very difficult	very low
57	0.57	0.33	4	2	average	good
58	0.76	0.5	6	3	easy	very good
59	0.24	0.5	3	0	difficult	very good
60	0.14	0	0	0	very difficult	very low
61	0.48	0.33	4	2	average	good
62	0.43	0.17	2	1	average	very low
63	0.48	0	3	3	average	very low
64	0.86	0.5	6	3	very easy	very good
65	0.86	0.17	6	5	very easy	very low
66	1	0	6	6	very easy	very low
67	0.62	-0.17	3	4	average	very low
68	0.52	0.33	4	2	average	good
69	0.67	0.33	5	3	easy	good
70	0.71	0.83	6	1	easy	very good
71	0.71	0.5	6	3	easy	very good
72	0.67	-0.17	3	4	easy	very low
73	0.43	0.33	3	1	average	good
74	0.14	-0.17	0	1	very difficult	very low
75	0.52	0.67	4	0	average	very good
76	0.52	0.33	4	2	average	good
77	0.9	0.17	6	5	very easy	very low
78	0.48	0.5	4	1	average	very good
79	0.67	0.83	6	1	easy	very good
80	0.86	0.17	6	5	very easy	very low

Item difficulty index was less than 0.60 and item discrimination index was less than 0.20 ($p < 0.60$ and $r < 0.20$) (except 25 questions which were found to be potentially problematic at the first stage of item analysis) these items are considered as non-discriminating items, it is evaluated that items 12, 13, 15, 27, 31, 36, 39, 48, 56, 60, 62, 63 and 74 cannot be certainly used in the finalized test form.

The expert opinion was determined to change the place of the 37th question from “Operating Systems” to “History of Computers”. Experts stated that the problem can be evaluated within the context of another sub-subject. Also, in consideration of the difficulty of the items in Table 3 and the criteria of item discrimination in Table 4, the number of questions that are not considered appropriate had increased to 31. Accordingly, the questions with the appropriate values were selected for each sub-subject and a list of 40 questions in Table 8 was obtained.

Table 8. Finalized Test Items for Achievement Test based on Content Tree

Week	Subject	Sub Subject	Count	Item Number
1	Excel I	What is Excel?	1	7
		Excel Layout	1	1
		Excel Home Tab	2	6, 8
2	Excel II	Excel File Tab	1	2
		Data Tab	3	11, 16, 17
		Insert Tab	2	14, 18
3	Excel III	Cells & Cell Range	1	22
		Arithmetic Formulas	1	21
		Arithmetic Operators	1	26
		Conditonal Formulas	2	29, 30
4	Computers & OS	How is it Works?	2	33, 34
		History of Computers	2	37*, 40
		Operating Systems	2	38
5	Internet & Web	Internet	1	49
		Web Developing Process	1	47
		Web Versions	1	42
		Web 2.0 Tools	2	44, 46
6	Cloud Systems	Cloud Technology	2	51, 58
		Cloud Technology Tools	3	53, 54, 57
7	CBL	Computer Based Learning	2	64, 68
		Instructional Softwares	1	67
		Types of Instructional Softwares	4	61, 69
8	Distance Education	Distance Education	2	71
		Advantages and Disadvantages	4	75, 76
		Tools	4	78, 79

Table 8 shows that included the questions and numbers for each subject. After excluding half of the test finalizing the test item statistics were changed positively. Mean item difficulty was found to be 0.638 as a result of the item analysis, and the difficulty of test items is average. Mean item distinctiveness is calculated to be 0.444 and the distinctiveness strength of the test items were very good acceptable.

4. Conclusion and Suggestion

Candidate teachers must have qualified education in ICT. Measurement and evaluation processes are conducted to determine the adequacy of this instruction. Here, the multiple-choice test which is widely used in different learning levels and groups (Çakan, 2017) has been preferred in conducting the measurement and evaluation processes.

Analyzes were conducted for validity and reliability, which are the two main components of the test (Büyüköztürk et al., 2015). To ensure validity, the content validity index was calculated as 0.91 with the opinions obtained from 8 field experts. This value indicates that content validity is achieved.

Test statistics were performed on 80 questions that provided content validity and 31 questions were found to be unsuitable for item difficulty and item distinctiveness values. Furthermore, 40 questions were obtained excluding the rest from the same sub-subject, so the target at the beginning of the test was achieved by choosing 40 questions.

Test statistics were calculated for 40 items as finalized test. The mean item difficulty value was 0.638, indicating that the test was of average difficulty. The average distinctiveness value is 0.444, which means that the test has very good distinctiveness. The reliability value was calculated with the KR-20 statistic and found to be 0.858. A value above 0.70 indicates that it is appropriate.

As a result of this unique study, a valid and reliable multiple-choice test consisting of 40 questions had expected

level of the difficulty and distinctiveness. It was provided a new assessment tool for the Informatics Education. It is believed that this assessment tool can help to identify the level of information and communication technologies of candidate teachers. Moreover, it is thought that it can be used by researchers not only working experimentally but also theoretically.

As a result of the findings obtained can make the following suggestions:

- The achievement test can be used to determine the adequacy of students' informatics skills in different informatics courses.
- By using this achievement test, feedback can be provided in students' misleading learning and misconceptions about informatics.
- This test can be used to determine the success level of the students in different researches.

5. References

Ayhan, İ. (2010). Eğitimcilere Yol Göstermesi Açısından Tab Analiz Programı Kullanarak Başarı Testi Hazırlama Sürecinde İzlenecek Adımlar. *Gümüşhane Üniversitesi Sosyal Bilimler Enstitüsü Elektronik Dergisi, 1*(2).

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

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

Büyüköztürk, Ş., Akgün, Ö. E., Demirel, F., Karadeniz, Ş., & Çakmak, E. K. (2015). *Bilimsel araştırma yöntemleri*: Pegem Akademi.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17): Sage publications.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*: ERIC.

Çakan, M. (2017). Eğitim sistemimizde yaygın olarak kullanılan sınav türleri. *Pegem Atıf İndeksi*, 87-122.

Eng, T. S. (2005). The impact of ICT on learning: A review of research. International Education Journal, 6(5), 635-650.

Garrett, H. E. (1937). Statistics in psychology and education.

Karasar, N. (2005). Bilimsel araştırma yöntemleri. *Ankara: Nobel Yayın Dağıtım, 15*.

Kozma, R. B. (2005). Monitoring and evaluation of ICT for education impact: A review. Monitoring and Evaluation of ICT in Education Projects, 19.

Lawshe, C. H. (1975). A quantitative approach to content validity 1. *Personnel psychology, 28*(4), 563-575.

McDougall, A., & Jones, A. (2006). Theory and history, questions and methodology: current and future issues in research into ICT in education. *Technology, Pedagogy and Education, 15*(3), 353-360.

Oliver, R. (2002). The role of ICT in higher education for the 21st century: ICT as a change agent for education. *Retrieved April, 14*, 2007.

Özçelik, D. A. (1992). *Ölçme ve Değerlendirme*. Ankara: ÖSYM Yayınları.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social work research, 27*(2), 94-104.

Sönmez, V., & Alacapınar, G. (2016). *Sosyal bilimlerde ölçme aracı hazırlama*: Anı Yayıncılık.

Sözbilir, M. (2010). Madde analizi ve test geliştirme. *Content Analysis and Test Development]. Alıntı Tarihi, 16*, 2013.

Şencan, H. (2005). *Güvenilirlik ve geçerlilik*: Hüner Şencan.

Toro, U., & Joshi, M. (2012). ICT in higher education: Review of literature from the period 2004-2011. International Journal of Innovation, Management and Technology, 3(1), 20-23.

Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science education, 16*(1), 199-207.

Veneziano, L., & Hooper, J. (1997). A method for quantifying content validity of health-related questionnaires. *American Journal of Health Behavior, 21*(1), 67-70.

Verhelst, N. (2014). Test Theory: Some Basic Notions Test Teorisi: Bazı Temel Kavramlar. *Education and Science, 39*(72), 3-19.

Yaşar, M. (2017). Ölçme ve değerlendirme ile ilgili temel kavramlar. *Pegem Atıf İndeksi*, 9-40.

Yıldırım, C. (1999). Eğitimde Ölçme ve Değerlendirme. *ÖSYM Yayınları, 4*.

Yurdugül, H. (2005). Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması. *XIV. Ulusal Eğitim Bilimleri Kongresi, 1*, 771-774.