Journal of Social Sciences and Education

Development and Validation of an Achievement Test in Biology

Hakkı ÇAKIR[1]
Yildiz Technical University, Graduate School of Science and Engineering, Science Education

Aslı GÖRGÜLÜ ARI[2]
Yildiz Technical University, Education Faculty, Department of Science and Mathematics Education

ABSTRACT

In this research, it was aimed to develop a valid and reliable test consisting of multiple choice questions in order to measure the conceptual understanding of the "cell" unit, which is one of the biology course topics. Planning, item writing, item analysis and item selection were followed in the development of the achievement test. Within the scope of the biology course, an item pool of 29 questions covering the cell topic was created. With the expert opinions, the content validity of the questions was ensured and the questions were corrected. Item analyzes were made with the pilot application of the draft form. As a result of item analysis, the number of questions was reduced to 22. After the item analyzes carried out, the "Cell Achievement Test", consisting of a total of 22 questions for the "cell" unit, took its final form. While the average item difficulty of the test was calculated as 0.55, the average item discrimination was calculated as 0.44. Moreover, the test was applied to a different group of 123 undergraduate students for reliability analysis, and the KR-20 reliability coefficient was calculated as 0.89. As a result of the findings, a valid and reliable "Cell Achievement Test" was obtained.

Key Words: Developing achievement test, biology, university course, validity, reliability.

Reserch Article

Received: 13.04.2022
Revision received: 24.05.2022
Accepted: 30.05.2022
Published online: 30.05.2022

[1] Corresponding author:
PhD Student
hakkickr@gmail.com
Orcid: 0000-0001-5732-8056

[2] *Assoc. Dr.*
agorgulu@yildiz.edu.tr
Orcid: 0000-0002-6034-3684

Introduction

The most generally accepted definition of education in our country (Arslan, Aydoğan, Ersözlü, İskender, Helvacı, & Turhan, 2009) is expressed as "the process of creating a desired change in an individual's behavior through his/her own life and intentionally" (Ertürk, 1975). A thriving education is possible if the curriculum includes strong relations between the four main dimensions of the curriculum, objective, content, learning-teaching process involved and assessment procedures (Demirel, 2005). "Measurament-Assessment process" is very important in determining the efficiency of teaching methods and techniques and the difficulties experienced by learners and their success in the learning process (Grounlund, 1985). Measurament-assessment is one of the most important components in the teaching and learning process. The quality and authenticity of measurement-assessment applications depend on valid and reliable test items. On the other hand, Turgut (1977) defined measurement as observing a quality and showing the observation result with numbers or other symbols. The traditional measurement tools that educators use during the process are: written exams, short-answer exams, true-false type tests, multiple-choice tests, matched tests, oral exams, and assignments (Karip, 2012; Yüksel, 2015). The most important reason for choosing multiple-choice questions from these traditional measurement tools is that they are easy to apply and have high content validity (Yaman, 2016). Although multiple choice questions are not sufficient to determine students' critical thinking skills and creativity, they are a preferred measurement tool to determine students' success and misconceptions (Küçükahmet, 2002). In addition, the advantages of multiple-choice exams are that the application time is short, they contain more questions, thus they are more valid and reliable, the scoring is objective and easy, the education can be applied from the first year to post- university, they can be applied to a large number of people at the same time, and they have statistically rapid calculation of accuracy and difficulty (Crocker & Algina, 2008).

Multiple choice tests are a measurement tool used for achievement tests. In this respect, achievement tests are one of the maximum proficiency tests to determine how much individuals have learned in the education process (Tezbaşaran, 2008). Achievement tests are important in terms of determining the meaningful learning of students, observing, analyzing and evaluating the change in students (Özcan, Çetinkaya, & Arık, 2021). On the other hand, multiple-choice tests are very suitable measurement tools to determine the knowledge level of a large number of students at different academic levels on different subjects (Burton, Sudweeks, Merrill, & Wood, 1991). When exam scores become the most important factor determining who is included and excluded from educational opportunities, scores that accurately reflect students' knowledge and skills become mandatory (Taylor & Walton, 1997).

When the literature is examined, in recent years, it is possible to encounter achievement tests prepared in the field of biology (Atik, Ekemen & Erkoç, 2019; Aymen-Peker & Taş, 2019; Güneş & Serdaroğlu, 2018; Güven, 2013; Kargın & Gül, 2021; Karslı et al., 2019; Nacaroğlu, Bektaş & Kızkapan, 2020; Şener & Taş, 2017; Şentürk & Selvi, 2021). In Table 1, achievement test studies conducted in the field of biology in recent years are given.

Table 1. *Achievement Test Studies in Biology Course*

Researcher	Research content	Sample Number (N)	Number of Items (n)	The Average Item Difficulty (pjx)	The Average Item Discrimination (rjx)	KR-20 Reliability Coefficient
Karslı,	Cell Divisions	409	36	0,48	0,45	0,86

Karamustafaoğlu & Kurt (2019)						
Aymen-Peker & Taş (2019)	Effect Wonder and Recognize World of Living Beings	210	23	0,47	0,52	0,80
Şentürk & Selvi (2021)	Human and Environment	273	27	0,62	0,47	0,82
Kargın & Gül (2021)	Body Systems and Health	390	40	0,61	0,47	0,86
Nacaroğlu, Bektaş & Kızkapan (2020)	Matter cycles and environmental problems	251	32	0,55	>0,19	0,81
Güven (2013)	Environmental problems	203	55	0,49	>0,21	0,87
Güneş & Serdaroğlu (2018)	Reproduction, Growth and Development in Plants and Animals	170	40	>0,36	>0,22	0,89
Kara-Ekemen, Atik & Erkoç (2019)	Biological Diversity	109	36	0,58	0,43	0,87
Şener & Taş (2017)	Systems in our Body	178	46	0,52	0,44	0,62
Adonu, Nwagbo, Ugwuanyi & Okeke (2021)	Supporting tissues in animals, axial skeleton, appendicular skeleton and joints	79	40	-	-	0,89

Pre-service science teachers have to learn courses such as physics, chemistry and biology at the university. It would be beneficial to teach the abstract concepts mentioned in the biology course by embodying them and using a constructivist approach (Öztürk & Karatekin, 2012). For researchers who adopt this approach or want to study in different types of approaches, students' achievements should be known. In this context, there is a need for cell success test, which is one of the basic subjects of biology at the undergraduate level, in order to determine the success of the new methods and approaches of the researchers. When the literature is examined, it is seen that many studies have been conducted on teacher candidates. A valid and reliable achievement test is needed in the studies to be conducted on the extent to which the applications made with the pre-service teachers within the scope of the biology course will affect the success of the pre-service teachers. Moreover, there is a need for an achievement test that can be used to determine the pedagogical content knowledge of biology teachers about cell. Furthermore, since the content of this achievement test is a "cell" subject and the same subject is included in high schools, it will also allow researchers who want to study with secondary education to benefit. The achievement tests in the literature on the subject of "cell" at university and high school levels were examined by the researcher, and no valid and reliable achievement test was found on the subject of "cell".

In this research, it was aimed to develop a valid and reliable test consisting of multiple choice questions in order to measure the conceptual understanding of the "cell" unit, which is one of the biology course topics.

- Is the achievement test consisting of multiple-choice questions prepared to measure the conceptual understanding of university students about the "cell" unit, a biology lesson, valid?
- Is the achievement test consisting of multiple-choice questions prepared to measure the conceptual understanding of university students about the "cell" unit, a biology lesson, reliable?

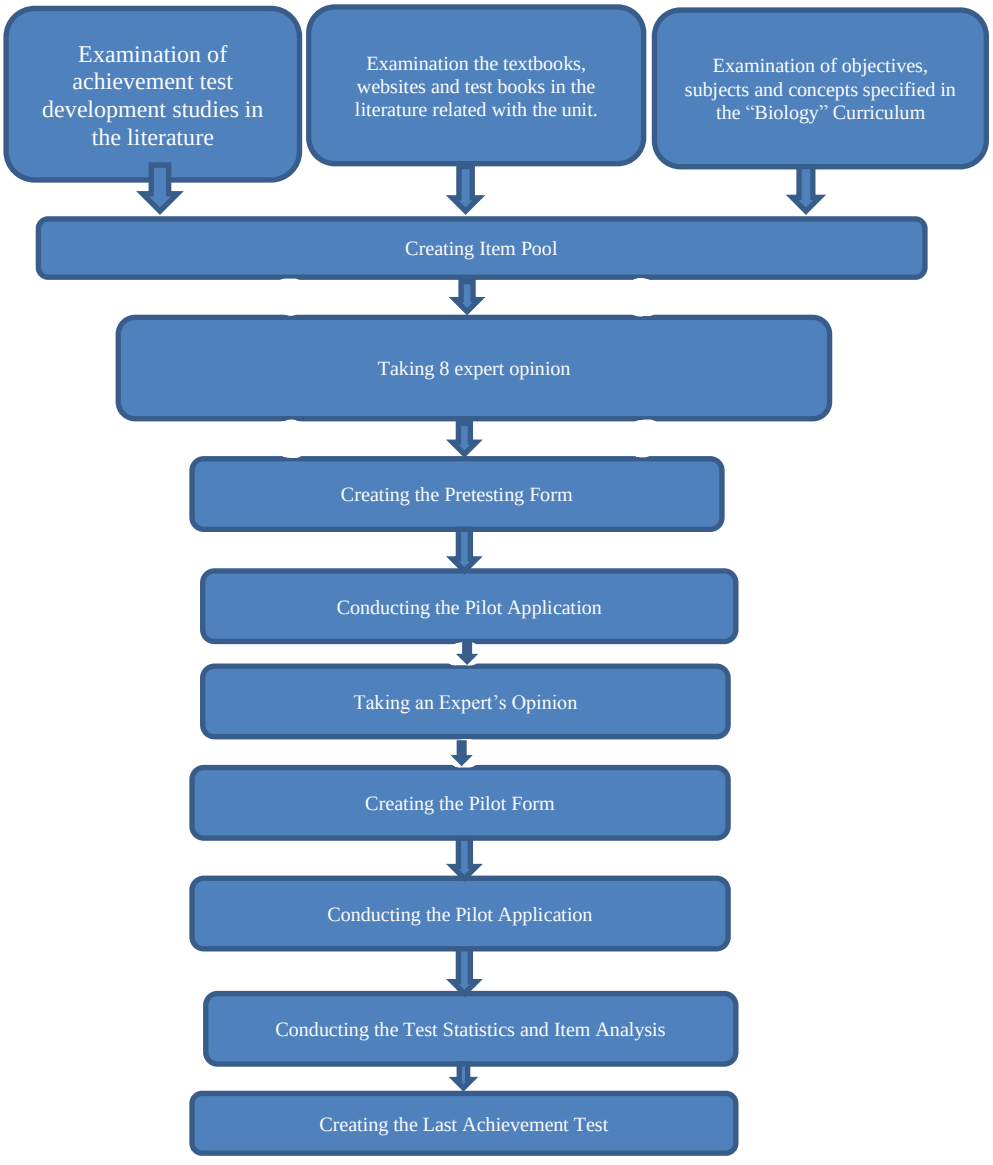
Method

Model

In this research, a valid and reliable test development study consisting of multiple-choice questions was conducted to measure the conceptual understanding of university students about the "cell" unit, which is one of the biology course topics.

The following stages were followed in the development of "Cell" Achievement Test (CAT). These stages are Çelik's (2000) academic achievement test development stages. These stages are planning, product writing, item analysis, and item selection, respectively. The figure developed by Şener & Öztürk (2017) was used and the figure-1 was created by editing. The figure developed by Şener & Öztürk (2017) was used and the figure-1 was created by editing.

Figure 1. *The process of developing achievement test (Şener & Taş, 2017).*



Sample and Population

In this research, it was aimed to measure the conceptual understanding levels of university students regarding the "cell" unit of the general biology course. 2nd grade science teacher candidates who studied at state universities and took biology courses participated in the research. Therefore, the sample of the research was determined according to the "criterion sampling" method. It is one of the sampling methods for criterion sampling and it is the creation of the sample from people, events, objects or situations with the qualities determined in the research (Gay, Mills, & Airasian, 2009).

The Development Process of the Achievement Test

The following stages were followed in the development of "Cell" Achievement Test (CAT). These stages are Çelik's (2000) academic achievement test development stages. These stages are planning, product writing, item analysis, and item selection, respectively.

Planning and Item Writing

First of all, it was determined what the undergraduate biology course curriculum was and from which sources the course was taught. Curriculum was examined. It has been determined that the Biology book of Campbell and Reece (2006) is widely used in universities as a biology textbook. Biology question sources were searched. The teaching materials of the general biology course and the course presentations created from the "Biology" book (Tepe, 2015) were accessed from foreign sources. This information obtained as a result of scanning formed the scope of this course. While preparing the academic achievement test item pool, the information in this survey was taken into consideration. A pool of 29 multiple choice questions was prepared by the researcher. In terms of content validity, the questions in the question pool are such as to cover the general information and concepts of the "cell" unit in the Biology book (Campbell & Reece, 2006). Moreover, it consists of a multiple choice item, a root, and a set of answer choices and it is more usual to offer four or five options (Anderson & Morgan, 2008).

Before the draft achievement test was finalized, 2 biology field experts, 1 biology education specialist, 2 biology teachers and 1 linguist examined the questions for content validity in order to control the language, content, writing style and page layout. After the examination, a 29-item draft achievement test was prepared with the necessary arrangements on the questions, taking into account the suggestions and criticisms from the experts. This draft achievement test was applied to 158 students as a pilot study. With the results obtained, item analyzes were made and 2 questions with item discrimination power below 0.19, 3 questions with an item difficulty index less than 0.80 and 2 questions with an item difficulty index less than 0.30 were excluded from the test. Since the item discrimination index of the items in the Cell Achievement Test was 0.30 and above, 2 questions were excluded from the test. In this state, the opinions of 2 biology experts were taken again and an achievement test was created.

The prepared achievement test was applied to 123 students who studied science teaching and took the "general biology" course in 2 different state universities in order to conduct reliability analysis.

Findings

Item Analysis and Item Selection

According to Smith (1991), validity is defined as the degree to which the researcher measures what he sets out to measure. Statistical operations can be performed with the "Simple Method" and the "Henrysson Method" for item analysis. While the Simple Method is

preferred in studies where the sample size is between 100-200 people, the Henrysson method is applied in cases where the sample size is less (Kuzu, 2008). While all participants are included in the calculation in the Henrysson method, 27% of the most successful and unsuccessful participants participate in the calculation in the Simple method (Turgut & Baykul, 2015).

Since the number of participants in the sample group was more than 100 in the first application (pilot application) and the last application, all data were included in the item analysis, that is, the Simple Method (27% lower and upper groups) was used at this stage.

The answers received from 158 students, to whom the draft achievement test was applied, formed the data of the item analysis. In the analysis, the point values of the questions belonging to the achievement test were determined as "1" point for each question. As a result of the item analysis, the item difficulty and item discrimination indices of each test question were calculated. Because the validity of the developed test, as well as its distinctiveness and suitability for the student's level, are revealed by the item difficulty index and item discrimination index (Reckase, 1985).

If the item difficulty index found as a result of the item analysis is close to 0, it means that the question is difficult, if it is close to 1, the problem is easy, and if it is between 0.40-0.60, it means that the question is of medium difficulty (Atılgan, 2009).

The item discrimination index for a question takes a value between +1 and -1. The negative value of the questions means that more people in the subgroup solved the question. If it is 0.40 and above, the item has a high discrimination power, if it is between 0.30 and 0.39, it is moderate, if it is between 0.20 and 0.29, the item discrimination is low, that is, the item needs to be corrected. The result is that it should be removed (Tekin, 2003).

While developing a multiple choice achievement test, the reliability of the scores needs to be examined. In cases where the difficulty indexes of the questions are close to each other, the reliability coefficient of Kuder Richardson-21 (KR-21) is used, while the reliability coefficient of Kuder Richardson-20 (KR-20) is used when the item difficulty indexes are not close to each other (Büyüköztürk, 2011). As a result of the item analysis, the item difficulty indexes were not equal to each other, so the K-20 calculation was made. The internal consistency of this achievement test was determined with the reliability coefficient KR-20 (Kuder Richardson-20). A general indicator of test quality is the KR-20, which is the coefficient of reliability that usually results from test analysis. It reflects the extent to which it will rank the same test takers when re-administered without any effect after the initial administration, in other words, the validity or discriminating power of the test. Values as low as 0.5 should give KR-20 values of 0.8 or higher (a maximum of 1.0) for tests containing more than 50 items, while satisfactory for short tests (10 to 15 items). Insufficiently low KR-20s are usually caused by an excess of too easy or difficult items, poorly written items that do not discriminate, or items that violate the prerequisite for testing a high-level content (Kehoe, 1994; Büyüköztürk, 2011).

Table 2. *Item Difficulty and Discrimination Index Values of the Questions in the Cell Achievement Test*

Item No*	**Item No****	**Item difficulty (pjx)**	**Item discrimination (rjx)**
Question 1	Question 1	0,61	0,51
Question 2	Question 2	0,55	0,40
Question 3	Question 3	0,69	0,47
Question 4	Excluded	**0,19**	0,31

Question 5	Question 4	0,52	0,45
Question 6	Excluded	**0,87**	0,36
Question 7	Question 5	0,61	0,56
Question 8	Question 6	0,50	0,44
Question 9	Question 7	0,66	0,45
Question 10	Question 8	0,53	0,41
Question 11	Excluded	0,61	**0,21**
Question 12	Question 9	0,55	0,39
Question 13	Question 10	0,30	0,33
Question 14	Question 11	0,56	0,41
Question 15	Excluded	0,41	**0,25**
Question 16	Question 12	0,55	0,39
Question 17	Question 13	0,51	0,37
Question 18	Question 14	0,63	0,35
Question 19	Excluded	**0,17**	0,34
Question 20	Question 15	0,75	0,52
Question 21	Excluded	**0,91**	0,45
Question 22	Question 16	0,53	0,42
Question 23	Question 17	0,51	0,46
Question 24	Question 18	0,52	0,50
Question 25	Question 19	0,59	0,49
Question 26	Excluded	**0,86**	0,35
Question 27	Question 20	0,56	0,44
Question 28	Question 21	0,59	0,48
Question 29	Question 22	0,57	0,50

*The item numbers of pilot test

** The item numbers of last test

This plot achievement test was applied to 158 students as a pilot study. As seen in Table 2, with the results obtained, item analyzes were made and 3 questions with an item difficulty index less than 0.80 and 2 questions with an item difficulty index less than 0.30 were excluded from the test. The average item difficulty index of the Cell Achievement Test, consisting of 22 questions, was calculated as 0.55.

As seen in Table 2, with the results obtained, item analyzes were made 2 questions with item discrimination power below 0.30. Since the item discrimination index of the items in the Cell Achievement Test was 0.30 and above, 2 questions were excluded from the test. The average item discrimination index of the Cell Achievement Test, consisting of 22 questions, was calculated as 0.44.

Moreover, the test was applied to a different group of 123 undergraduate students for reliability analysis, and the KR-20 reliability coefficient was calculated as 0.89.

Discussion and Result

In this research, the "Cell Achievement Test", which includes 22 multiple-choice questions, was developed for undergraduate students about the basic unit of life, the cell, and the structure and function of the cell. In the research, the difficulty index (Pj), item

discrimination index (rjx) and KR-20 Reliability Coefficient of each item were calculated through the item analysis performed within the validity.

The item difficulty index, which shows the correct answer rate for each item in the prepared test, takes values ranging from "0" to "1", and if this value is close to 0, the item is said to be difficult, and if it is close to 1, the item is said to be easy. Especially, this value is expected to be between 0.20 and 0.80 in achievement tests (İlhan & Hoşgören, 2017). The difficulty index of the 4th and 19th items in the Cell Achievement Test is below 0.20; Since the difficulty indexes of the 6th, 21st and 26th items were above 0.80, it was decided to exclude these items from the test. As such, as stated in the Table 2, the difficulty index values of the remaining items in the test vary between 0.30 and 0.75.

The item discrimination index value, on the other hand, takes values ranging from "-1" to "+1" as stated in the literature. If this value is close to zero, the item is insufficient to distinguish the upper and lower groups; Approaching the +1 value indicates that the item has a high distinctiveness value. In addition, if the item discrimination index is negative, it is interpreted that the relevant item is answered more by those in the subgroup, and in this case, the prepared test does not serve its purpose adequately (Kubiszyn & Borich, 2003). If the discrimination index of an item selected in a prepared test is 0.19 and below this value, it should be removed from the test because it is a very weak item; If a value between 0.20-0.29 is found, the item should be corrected and improved; 0.30-0.39 is pretty good but still can be improved; If it is 0.40 and above, it is interpreted as an item with very good discrimination (Karslı & Ayaz, 2013; Tosun & Taşkesenligil, 2011). The discrimination index value of the 11th and 15th items in the Cell Achievement Test is below 0.30 and it was decided to exclude these items from the test. As such, as stated in the Table 2, the difficulty index values of the remaining items in the achievement test developed are above 0.30.

Before the pilot achievement test was finalized, 2 biology field experts, 1 biology education specialist, 2 biology teachers and 1 linguist examined the questions for content validity in order to control the language, content, writing style and page layout. After the examination, a 29-item draft achievement test was prepared with the necessary arrangements on the questions, taking into account the suggestions and criticisms from the experts. This draft achievement test was applied to 158 students as a pilot study. With the results obtained, item analyzes were made and 2 questions with item discrimination power below 0.30 since the item discrimination index of the items in the Cell Achievement Test was 0.30 and above. 3 questions with an item difficulty index less than 0.80 and 2 questions with an item difficulty index less than 0.30 were excluded from the test so, the difficulty index values of the remaining items in the test vary between 0.30 and 0.75. In this state, the opinions of 2 biology experts were taken again and an achievement test was created. This process contributed to increase the validity of this achievement test.

The prepared achievement test was applied to 123 students who studied science teaching and took the "biology" course in 2 different state universities in order to conduct reliability analysis. The KR-20 reliability coefficient was calculated as 0.89. These values have shown that this achievement test is reliable.

In conclusion, it can be declared that this test, which was developed to determine the achievement levels of undergraduate students in the cell subject, has reliable results in terms of validity and reliability. In addition, it can be stated that the difficulty and discrimination levels of the items in the test meet the desired criteria and have the ability to distinguish academically low and high students.

As a result of the findings, a valid and reliable "Cell Achievement Test" was obtained.

Recommendations

After this test is applied in secondary education, item analyzes can be done. For high school students, a valid and reliable achievement test can be created quickly by using the items in this test. In addition, an attitude scale can be developed to determine the attitudes of university students towards the subject of "cell".

Acknowledgment

This research was produced from first author's PhD thesis and supported by Scientific Research Project Number SBA-2021-4253 within the University Project of Supporting Program for Normal Research Project in Yildiz Technical University.

References

Adonu, C. J., Nwagbo, C. R., Ugwuanyi, C. S., & Okeke, I. O. (2021). Improving Students' Achievement and Retention in Biology using Flipped Classroom and Powerpoint Instructional Approaches: Implication for Physics Teaching. *International Journal of Psychosocial Rehabilitation,* 25(2), 234-247.

Anderson, P., & Morgan, G. (2008). National assessments of educational achievement, 2: Developing tests and questionnaires for a national assessment of educational achievement. World Bank Publications.Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412.

Arslan, M., Ersözlü, Z. N., Aydoğan, İ., İskender, M., Helvacı, M. A., & Turhan, M. (2009). Eğitimle ilgili temel kavramlar. In Arslan (Eds.), Eğitim Bilimine Giriş (12-25). Ankara: Gündüz.

Atik, A. D., Ekemen, D. K., & Erkoç, F. (2019). "Biyolojik çeşitlilik ve korunması" konusunun öğretilmesinde istasyon tekniği kullanımının öğrencilerin başarısı üzerine etkisi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi,* 19(4), 1499-1513.

Atılgan, H. (Ed.) (2009). Eğitimde ölçme ve değerlendirme (4. Baskı). Ankara: Anı.

Aymen Peker, E. & Taş, E. (2019). Üç Aşamalı Kavram Tanı Testi Geliştirme Çalışması: 5. Sınıf Canlılar Dünyasını Gezelim Tanıyalım Ünitesi. *Journal of Computer and Education Research,* 7 (14), 515-539. DOI: 10.18009/jcer.602223

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Brighaam young university testing.

Büyüköztürk, Ş. (2011). Sosyal bilimler için veri analizi el kitabı. Ankara. Pegem Akademi.

Campbell, Neil A., Reece, Jane B. (2006). Biyoloji (6). (E. Gündüz, A. Demirsoy, İ. Türkan, Çev.). Ankara: Palme.

Çelik, D. (2000). Okullarda Ölçme Değerlendirme Nasıl Olmalı? (1. Baskı). Ankara: MEB.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Mason, Ohio: Cengage Learning

Demir, B. ve Akarsu, N. (2014). Modern Fizik Konuları ile İlgili Kavram Testi Geliştirilmesi ve Uygulanması: Modern Fizik Kavram Testi (MKFT). *Journal Of European Education*, 4(2).

Gay, L. R., Mills, G. E., & Airasian, P. W. (2009). Educational research: Competencies for analysis and applications. Merrill/Pearson.

Gönen, S., Kocakaya, S. ve Kocakaya, F. (2011). Dinamik konusunda geçerliliği ve güvenilirliği sağlanmış bir başarı testi geliştirme çalışması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi,* VIII (I), 40-57.

Gronlund, N. (1985) Measurement and evaluation in teaching. New York: MacMillan.

Güneş, M. H. & Serdaroğlu, C. (2018). Bitki ve Hayvanlarda Üreme, Büyüme ve Gelişme Ünitesinde Geliştirilen Başarı Testinin Geçerliliği ve Güvenirliği. *Bilim Eğitim Sanat ve Teknoloji Dergisi,* E 2 (1), 34-40.

Güven, E. (2013). Çevre Questionnları Başarı Testinin Geliştirilmesi ve Öğretmen Adaylarının Bilgi Düzeylerinin Belirlenmesi. *Trakya Üniversitesi Eğitim Fakültesi Dergisi,* 3 (2).

Kara Ekemen, D., Atik, A. D. & Erkoç, F. (2019). "Biyolojik çeşitlilik ve korunması" konusunun öğretilmesinde istasyon tekniği kullanımının öğrencilerin başarısı üzerine etkisi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi,* 19 (4), 1499-1513. DOI: 10.17240/aibuefd.2019.-518071

Karatekin, P., ve Öztürk, M. (2012). Fen ve teknoloji öğretmen adaylarının genel biyoloji laboratuarında TGA tekniğiyle işlenmiş "Hücre ve Dokular" ünitesinin öğrencilerin başarı ve bilimsel süreç becerileri üzerine etkisi. *Manisa Celal Bayar Üniversitesi Eğitim Fakültesi Dergisi*, 2(1), 106-131.

Kargın, P. D. & Gül, Ş. (2021). Altinci sinif "vücudumuzdaki sistemler ve sağlığı" ünitesine yönelik bir başari testi geliştirilmesi. *Ihlara Eğitim Araştırmaları Dergisi*, 6 (1), 1-26. DOI: 10.47479/ihead.729412

Karip, E. (2012). Ölçme ve değerlendirme. Ankara: Pegem Akademi.

Karslı G, Karamustafaoğlu S, Kurt M (2019). Yenilenen Fen Bilimleri Dersi Öğretim Programına Yönelik 7. Sınıf "Hücre ve Bölünmeler" Ünitesi Başarı Testi: Geçerlik ve Güvenirlik. *Fen Bilimleri Öğretimi Dergisi*, 7(1), 68 - 98.

Kaya, S., İnaç, H., & Çelik, H. Assure öğretim tasarımı uygulamalarının öğrencilerin akademik başarısı üzerine etkisi. In International Marmara Sciences Congress (AUTUMN 2020) (p. 486).

Kehoe, J. (1994). Basic item analysis for multiple-choice tests. *Practical Assessment, Research, and Evaluation*, *4*(1), 10.

Kenan, O. ve Özmen, H. (2014). Maddenin tanecikli yapısına yönelik iki aşamalı çoktan seçmeli bir testin geliştirilmesi ve uygulanması. *Journal of Research in Education and Teaching*, 3(3), 371-378.

Küçükahmet, L. (Ed.) (2002). Sınıf yönetimi. Ankara: Nobel.

Kuzu, A. (2008) "Ölçme aracı geliştirme, test ve madde analizi", Bilimsel Araştırma Yöntemleri ve Ölçme Değerlendirme, A. Aşkım Kurt (Ed.), Anadolu Üniversitesi, Eskişehir

Nacaroğlu, O., Bektaş, O., & Kızkapan, O. (2020). Madde döngüleri ve çevre Questionnları konusunda başarı testi geliştirme: Geçerlik ve güvenirlik çalışması. *Kastamonu Eğitim Dergisi,* 28(1), 36-51.

Özcan, H., Çetinkaya, İ., & S., Arık. Ortaokul Öğrencilerinin Basit Makineler Ünitesi ile İlgili Anlayışlarını Ölçmeye Yönelik Bir Test Geliştirme Çalışması. *Muğla Sıtkı Koçman Üniversitesi Eğitim Fakültesi Dergisi*, *8*(1), 16-36.Tekin, H. (2003). Eğitimde Ölçme ve Değerlendirme. (16. baskı). Ankara: Yargı.

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

Şentürk, Ö. Ç., & Selvi, M. (2021). Fen Bilimleri Dersi "İnsan ve Çevre" Ünitesi Akademik Başarı Testi Geliştirme: Güvenirlik ve Geçerlik Çalışması. Gazi Üniversitesi *Gazi Eğitim Fakültesi Dergisi,* 41(2), 601-630.

Şişman, M. (2011). Eğitim bilimine giriş.

Taylor, K. & Walton, S. (1997). Co-opting standardized test in the service of learning. Phi Delta Kappan, 66-70.

Tepe, B. (2015). Hücre içinde yolculuk, hücre döngüsü, mayoz ve eşeyli yaşam döngüleri. [Powerpoint slides] http://bektastepe.net/course-slides/general-biology.html

Tezbaşaran, A. A. (2008). Likert tipi ölçek hazırlama kılavuzu. Ankara: Türk Psikologlar Derneği.

Turgut, F. ve Baykul, Y. (2015) Eğitimde Ölçme ve Değerlendirme. Ankara: Pegem Akademi.

Yaman, S. (2016). Çoktan Seçmeli Madde Tipleri ve Fen Eğitiminde Kullanılan Örnekleri. *Gazi Eğitim Bilimleri Dergisi*, *2*(2), 151-170.

Yüksel, M. (2015). *Öğrenci başarısının belirlenmesinde kullanılan farklı ölçme araçlarından elde edilen puanların karşılaştırılması* (Master's thesis), Gaziosmanpaşa Üniversitesi, Istanbul.