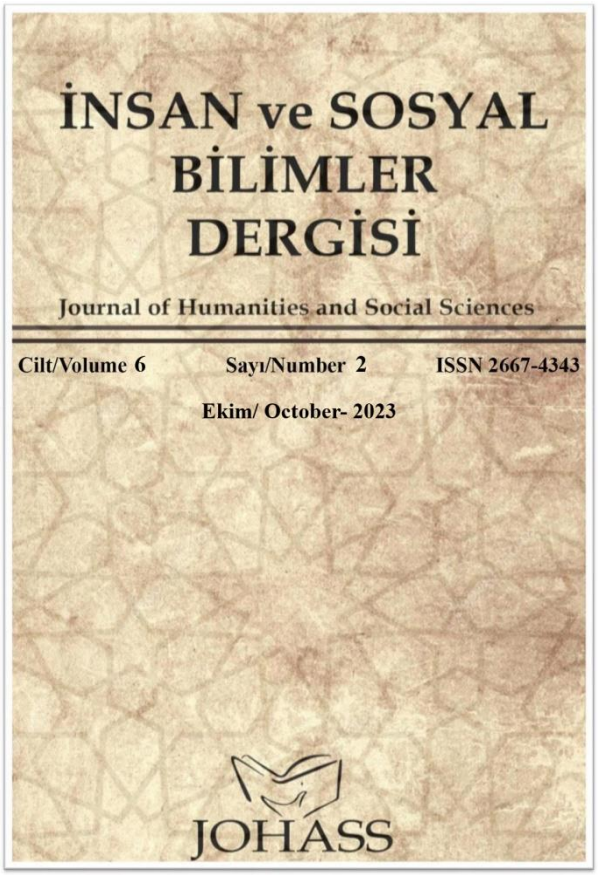
**JOURNAL OF HUMAN AND SOCIAL SCIENCES (JOHASS)**




**https://dergipark.org.tr/tr/pub/johass**

# A Review of Achievement Test Development in Türkiye Regarding the Achievement Test Development Process

**Müge ULUMAN**[1]
*Marmara Univerity, Atatürk Faculty of Education, Department of Educational Sciences*
*Assit. Prof. Dr.*
*mugeulumann@gmail.com*
*Orcid ID: 0000-0003-4155-3114*

**Article Type:** Research Article
Received: 30.09.2023
Revision received: 19.10.2023
Accepted: 26.10.2023
Published online: 27.10.2023

**Citation:** Uluman, M. (2023). A review of achievement test development in Türkiye regarding the achievement test development process. *Journal of Human and Social Sciences*, *6*(2), 377-405.

# A Review of Achievement Test Development in Türkiye Regarding the Achievement Test Development Process

**Müge ULUMAN**[1]

*Marmara Unıversıty, Atatürk Faculty of Educatıon, Department of Educatıonal Scıences*

**Abstract**

The purpose of this research is to determine to what extent the articles titled achievement test development, published between 2020-2023, meet the steps that should be followed during the achievement test development process. The research is a document review study within the scope of qualitative research. In determining the studies to be included in the research, consideration was given to the fact that the aim and title of the research included '...developing an achievement test', that the full text of the article was accessible, and that the research language was Turkish, and a total of 40 articles were included in the research. In the study, a coding list was created and used by the researcher, based on Turgut and Baykul's (2012) achievement test development steps. Data analysis of the research was carried out using the categorical analysis method, one of the types of content analysis. In the articles examined within the scope of the research, it is seen that there are deficiencies in almost all stages of the achievement test development process, starting from presenting the test plan based on the test plan to including the items to be included in the final test form. In line with the research results, it may be suggested that the achievement test development steps in the literature be taken into consideration in studies.

**Keywords:** Achievement test, achievement test development steps, document review

**Research Article**

*Received*: 30.09.2023
*Revision received:* 19.10.2023
*Accepted:* 26.10.2023
*Published online:* 27.10.2023

[1] Corresponding author:
*Assit. Prof. Dr.*
*mugeulumann@gmail.com*
*Orcid ID: 0000-0003-4155-3114*

## Introduction

To establish an effective education system, it is imperative to assess the efficacy of educational policies and curricula, the capabilities of the educators responsible for instruction, and the identification of potential challenges within the educational process. Among these factors, student achievement stands out as one of the most critical indicators for evaluating the functionality of the education system. Tests designed to gauge student achievement serve as the foundation for decisions related to teaching, guidance, administration, and research. In light of the significance of achievement tests in education, it is crucial to engage in a comprehensive discussion concerning their classifications, attributes, and the steps involved in their development

Various definitions emerge when delving into the literature on achievement tests, although they share common elements. These definitions encompass measurement tools specifically designed to assess the general knowledge acquired at school and in life (Heckman & Kautz, 2012). They consist of a series of questions aimed at evaluating an individual's learning outcomes after exposure to educational experiences (Anderson, 1972). Additionally, they measure the knowledge or skills acquired throughout a student's educational process (Popham, 2008). Some refer to them as maximum performance tests, focusing on identifying what an individual has learned (Thorndike & Thorndike-Christ, 2013), while others view them as tools to gauge the attainment of educational objectives and desired behaviors (Turgut & Baykul, 2012). However, it is also seen that there are different classifications of achievement tests and different names (standard, standard-based, national scale standard, commercial, international scale standard, criterion-based, norm-based, teacher-made achievement tests, and aptitude tests (Airasian, 2001; Brookhart & Nitko, 2019; Popham, 2008) in the literature. These classifications are based on what the test is intended to measure and whether it has the distinctive features of standardized tests. Since it is not the subject of this study, comprehensive information is not provided. However, detailed insights into standardized tests and their characteristics can be found in Koç's (1985) study.

In the creation of achievement tests, various item types with their distinct advantages and disadvantages are employed (multiple-choice, open-ended, short-answer, true-false, etc.). Among these item types, the most commonly used is the multiple-choice item due to its advantages (Fellenz, 2004; Saadat et al., 2021; Anderson, 2003; Aydın, 2018; Chatterji, 2003; Cheung, 2003; Haladayna, 2004; Kline; 2000; Kubiszyn & Borich, 1996; Miller et al., 2012;

Singh & Rosengrant, 2003; Özçelik, 2010). In order for an achievement test to be considered good, it is expected to have reliability, validity, and usefulness. These characteristics can only be achieved by following a certain systematic process during the development of an achievement test. The process referred to as the development of achievement tests, although containing similar points in the literature, has been divided into and labeled as stages in different ways by various researchers (Crocker & Algina, 1986; Haladyna, 1997; Irwing & Hughes, 2018; Lane et al., 2015; Porter, 2002; Turgut & Baykul, 2012; Webb, 1997). One of the most frequently used achievement test development processes in the literature belongs to Webb (1997) and the test development process consists of four criteria. The first criterion is the categorical concurrence criterion, which provides information about the extent to which the items in the test cover the objectives. In this criterion, items and objectives are matched by experts, and the percentage of concordance is examined. The second criterion is depth of knowledge consistency. In this criterion, what is expected from the students in the context of the objectives is revealed and items are developed. The percentages of concordance between the objectives and items are determined. In the third criterion, range of knowledge correspondence, it is necessary to determine the concordance between the behavior that students should have with the objectives and the behavior tested in the items. The last criterion, balance of representation, examines the distribution of items to objectives by calculating the balance index. Another one belongs to Turgut and Baykul (2012), and they have expressed these steps as follows: determining the purpose of the test, and consequently, the objectives for which the test scores will be used; listing the learning outcomes to be measured in the test (e.g., objectives); defining the subject area to be covered by the test; creating a specification table; linking the item type and items to the learning outcomes and the underlying taxonomy; determining the number of items and the duration; specifying and implementing the method to be followed in drafting and editing items and item selection; determining and implementing methods for test construction, replication, administration, and scoring.

The quality of the scores obtained from achievement tests can control efforts to improve education, guide instruction (Kimberlin & Winterstein, 2008), and strengthen decisions made about individuals. The development of achievement tests has been extensively studied by different researchers in different fields and age groups in the literature. These include science (Açıkgöz & Karslı, 2015; Akbulut & Çepni, 2013; Ayvacı & Durmuş, 2016; Bolat & Karamustafaoğlu, 2019; Armağan & Demir, 2019; Güven, 2013; Saraç, 2018; Şen &

Eryılmaz, 2011), mathematics (Akkuş & Akkaş, 2021; Ersoy & Bayraktar, 2018; Karaboğaz & Ergene, 2023; Narlı & Başer, 2008; Şahin et al., 2023). Achievement tests, which are so frequently studied and are very important in the educational decisions made, and the steps followed regarding the development processes and how they are reported are considered very important. There are some studies in the literature in which achievement test development processes are examined. In his study, Karadağ (2011) examined a total of 77 theses used in doctoral dissertations in educational sciences between 2003-2007 by taking into account the achievement test development processes. He evaluated these tools as having low quality for reasons such as not determining reliability and validity in achievement tests and not including information on how to score the test. Mutluer and Yandı (2012) examined 50 undergraduate theses published between 2010 and 2012 within the scope of educational sciences institutes and determined the theses that acted in accordance with the test development steps and those that did not. Boyraz (2018) examined the multiple-choice achievement tests used in doctoral dissertations developed within the scope of the Department of Elementary Education between 2012-2017. In this study, which deals with the achievement test development process holistically, it was concluded that the achievement tests used in the theses did not meet the desired quality standards. Şahin et al. (2023) examined a total of 39 achievement test development articles published in the journals they identified between 2015 and 2020 in the field of mathematics. In the majority of the articles, it was determined that the item pool was not prepared, pilot application was not carried out, item analysis was omitted, sample items were absent, and there were certain deficiencies in terms of validity. Considering these studies, it is important to examine the current situation regarding achievement test development studies without any field restriction. This study was conducted to discuss the achievement test development process in order to set an example for both researchers and practitioners. Therefore, the aim of the study is to determine to what extent the articles published between 2020-2023 with the title of achievement test development meet the steps to be followed during the achievement test development process.

## Method

### Model

This research examines the studies that develop achievement tests by taking into account the points to be considered when developing achievement tests. In this respect, the

research is a document review study within the scope of qualitative research. Document analysis involves the process of collecting data by examining existing records or documents (Yıldırım & Şimşek, 2013).

**Documents**

The aim of the research is to examine the studies in the national literature between 2020-2023, which can be accessed online. "When selecting studies to be included in the scope of this research, we considered whether the research had the purpose and title of '...developing an achievement test,' whether the full article text was accessible, and whether the research was conducted in the Turkish language. Studies that did not meet these criteria and thesis studies were not included in the study. A total of 40 articles were accessed and included in the study. Information about the studies included in the research is presented in Table 1.

**Tablo 1**

*Distribution of articles ın Research according to Some Variables*

| **Variables** | | **f** | **%** |
|---|---|---|---|
| **Year** | 2020 | 19 | 47.5 |
| | 2021 | 10 | 25 |
| | 2022 | 8 | 20 |
| | 2023 | 3 | 7.5 |
| **Test Developed Area** | Science (primary education level) | 3 | 7.5 |
| | Life Science | 1 | 2.5 |
| | Maths | 2 | 5 |
| | Science | 24 | 60 |
| | Turkish | 1 | 2.5 |
| | Biology | 1 | 2.5 |
| | T.R. The History of Revolution | 1 | 2.5 |
| | Nursing | 1 | 2.5 |
| | Digital Citizenship | 1 | 2.5 |
| | Informatics | 1 | 2.5 |
| | Early Literacy | 1 | 2.5 |
| | Basic Laboratory | 1 | 2.5 |
| | International Passenger Transportation | 1 | 2.5 |
| | Measurement and Evaluation Techniques | 1 | 2.5 |
| **Grade Level** | 3rd Class | 1 | 2.5 |
| | 4th grade | 3 | 7.5 |
| | 5th grade | 4 | 10 |
| | 6th grade | 8 | 20 |
| | 7th grade | 9 | 22.5 |
| | 8th grade | 7 | 17.5 |
| | 11th grade | 1 | 2.5 |
| | Undergraduate | 5 | 12.5 |
| | Adult | 1 | 2.5 |
| | Teachers | 1 | 2.5 |

Considering Table 1, it can be stated that the studies included in the research were published in 2020 at most and in 2023 at least, and that most of them were carried out in the field of science and secondary school grade levels.

**Data Collection Tools**

Upon reviewing the literature, it becomes evident that, despite the considerable similarity in the achievement test development process, various researchers have delineated its steps in different manners (Crocker & Algina, 1986; Haladyna, 1997; Irwing & Hughes, 2018; Lane et al., 2015; Porter, 2002; Turgut & Baykul, 2012; Webb, 1997). For this study, we relied on the achievement test development steps outlined by Turgut and Baykul (2012) because the steps were understandable and clear and access to the source was easy. To facilitate the examination of the selected articles, we devised a coding list centered on the research's objectives, the steps involved in test development, and considerations throughout the test development process. This coding list was presented to five field experts who completed their doctorate in the field of measurement and evaluation and their feedback was received. In line with this feedback, the list was revised and finalized. In the coding list, firstly, preliminary information about the articles was presented. Then, 36 items with statements graded as yes, partially and no were included.

**Collection of Data and Analysis**

The data analysis of the research conducted within the scope of qualitative research was carried out by using the categorical analysis method, which is one of the types of content analysis. Categorical analysis can be defined as dividing the subject or situation under investigation into units and grouping each unit into categories according to certain criteria (Mayring, 2004). There are two different ways of conducting categorical analysis: the Theoretical Categorization Process and the Applied Categorization Process. In the process of Theoretical Category Formation, categories can be determined initially as there is a much clearer theoretical foundation to start from. In the process of Applied Category Formation categories can be created as the relevant materials begin to be examined. If needed in the course of research, both approaches can be used simultaneously (Tavşancıl & Aslan, 2001). In this study, since the theoretical basis for the achievement test development process was relatively clear, the theoretical categorization process was employed. The coding process was

carried out by the researcher using a coding list and Microsoft Office Excel 2016 software. Then, frequency values were obtained for each item in the list according to the coding.

The reliability of the data obtained in the context of the categorical analysis method depends especially on the coding process. Ratings of categories, whether the researcher conducts coding at two different times or if the codings remain consistent across different researchers, provide reliability as an indicator of objectivity (Tavşancıl & Aslan, 2001). All articles within the scope of the research were examined by the researcher. Moreover, ten randomly selected articles were asked to be coded by a field expert who completed his/her doctorate in the field of measurement and evaluation in line with the coding list and the agreement between these two researchers was examined. As evidence of reliability in categorical analysis studies, the percentage of agreement between the researchers was calculated using the formula developed by Miles and Huberman (1994) (Reliability= number of agreement / (number of agreement + number of disagreement)). The reliability coefficient calculated based on this formula was found to be 0.91, which is considered as a high level (Cohen & Swerdlik, 2010). It can be stated that this value serves as evidence of the high reliability of the data obtained from the study. In addition, all documents examined within the scope of the research are included in the bibliography to contribute to the reliability of the research (Yin, 2014).

## Findings

In the study, 40 articles aiming to develop an achievement test were analyzed in line with the coding list developed by the researcher. The findings obtained in the context of the headings in the coding list are given in the tables below. First of all, Table 2 shows the distribution of the data obtained to determine the purpose of the test.

**Table 2**

*Frequency Values About Purpose of The Test*

| | **Yes** | | **Partial** | | **No/NI** | |
|---|---|---|---|---|---|---|
| | **f** | **%** | **f** | **%** | **f** | **%** |
| Test plan has been created. | 18 | 45 | 6 | 15 | 16 | 40 |
| The purpose of the test is clearly set out. | 26 | 65 | 11 | 27.5 | 3 | 7.5 |
| Information about the target audience for the test has | 34 | 85 | 6 | 15 | | |

been provided.

When considering Table 2, the initial examination focused on whether a test plan, which is not a formal step in the achievement test development process but provides guidance on systematic progression within this process, had been established. It is seen that 18 of the articles created a test plan and developed the achievement test in line with this plan. In six of the articles, the word test plan or achievement test development process was used, but these steps were not clearly and explicitly stated. Furthermore, 16 articles did not include a test plan or the steps of the achievement test development process. When examining whether the purpose of the test, regarded as the initial phase of achievement test development, is explicitly stated, it is possible to say that 26 of the articles have successfully implemented this stage. However, in 11 articles, it can be stated that the purpose was not clearly explained and the purpose of the achievement test was considered as the purpose of the article. It was evaluated as partial on the grounds that the purpose of the article did not fully meet the purpose of developing an achievement test. Determining which audience the developed test will serve is one of the important steps for the achievement test. In 34 of the articles, this step was present but it was not articulated clearly. Instead, it was evaluated as partial since the relevant information was extracted from the details provided under the research study group's title. Table 3 below shows the frequencies of the learning products to be tested with the items in the test and the characteristics of the test.

**Table 3**

*Frequency Values for The Characteristics of Learning Products and Achievement Test*

| | **Yes** | | **Partial** | | **No/NI** | |
|---|---|---|---|---|---|---|
| | **f** | **%** | **f** | **%** | **f** | **%** |
| All learning products are included. | 33 | 82.5 | 4 | 10 | 3 | 7.5 |
| The subject area in which the learning products are located is indicated. | 33 | 82.5 | | | 7 | 17.5 |
| The unit area in which the learning products are located is indicated. | 30 | 75 | | | 10 | 25 |
| A taxonomy was used as a basis for classifying learning products. | 25 | 62.5 | 3 | 7.5 | 12 | 30 |
| The number of items is included. | 40 | 100 | | | | |
| The duration of the test is indicated | 14 | 35 | | | 26 | 65 |
| The types of items used are indicated. | 39 | 97.5 | 1 | 2.5 | | |
| The number of options is given. | 26 | 65 | | | 14 | 35 |

Any achievement test is designed to measure specific learning products within a specific unit or subject area. Therefore, it is important to present this scope in a clear and understandable way in the achievement test development process. In this context, when it is examined whether the articles included all of the learning products that they aimed to measure, it can be stated that 33 of the articles did so. It is seen that four of the articles only defined the general scope of the relevant learning products but did not include what this scope includes, and three of them did not provide sufficient and comprehensible information by pointing to the number of learning products, grade level or the area in which the test was developed. Article 18, which was evaluated as no in this item, gave information as follows: '...related to the subject, the needs were determined by literature review, opinions from experts and students who conducted academic studies, and 23 outcomes emerged.' The article contained no additional information beyond this. According to Table 3, it can be stated that most of the articles contain information about the subject area and units in which these learning products are included. In the classification of learning products, 25 of the articles are based on a classification, with Bloom's classification being the most commonly used, while Haladayna's classification is used to a lesser extent (5). While 12 of the articles did not mention any classification at all, one of the articles mentioned the underlying classification in the introduction but did not include any information in the other sections where achievement test development was explained. In two articles, the steps of the underlying classification were included in the specification sheet, but no other information was provided. Therefore, these three articles were categorized as partial. All articles included the number of items in the final test, but the test duration was featured in 14 articles and omitted in 26. The type of item to be used in the achievement test was only mentioned in the introduction of one article and was evaluated as partial on the grounds that it was understood from this information what the type of item in the test was. However, it is seen that multiple-choice test items are used in all achievement tests. Information on how many answer options multiple-choice items have was included in 26 articles, while 14 articles did not provide this information. Information and frequency values related to content validity and creation of the item pool are presented in Table 4.

**Table 4**

*Frequency Values for Content Validity and İtem Pooling*

| | **Yes** | | **Partial** | | **No/NI** | |
|---|---|---|---|---|---|---|
| | **f** | **%** | **f** | **%** | **f** | **%** |
| Specification table was prepared. | 30 | 75 | 3 | 7.5 | 7 | 17.5 |
| Expert opinion was taken for the prepared specification table. | 18 | 45 | | | 22 | 55 |
| An item pool was created in line with the specification table. | 30 | 75 | | | 10 | 25 |
| The rules taken into consideration while writing the trial items were included. | 6 | 15 | 8 | 20 | 26 | 65 |
| Expert opinion was taken for the written items. | 38 | 95 | | | 2 | 5 |
| Any statistical analysis was used to determine content validity. | 8 | 20 | | | 32 | 80 |
| Face validity was given. | 14 | 35 | | | 26 | 65 |

It is possible to say that specification tables have a very important role in achievement tests, especially to ensure content validity. When examining whether the specification table was prepared in the achievement tests, it was determined that the specification table was included in 30 of them, while it was not provided in seven of them. In three articles, it was found that the information that the table of specifications was prepared was given but this table was not included. While expert opinion was taken in 18 articles for the tables of specifications prepared, it was determined that expert opinion was not taken in 22 articles. The specification table prepared in the writing stage of the items to be included in the trial form of the achievement test is taken as a basis. When Table 4 is taken into consideration and in parallel with the item "a specification table was created", 30 of the articles acted in this direction, while 10 of the articles did not include any information or explanation. Various writing rules should be taken into consideration in the context of the type of item used in the writing phase of the items to be included in the trial test form. Since multiple-choice items were used in all of the articles examined within the scope of this study, there are rules to be considered in the writing of this item type in the literature (Haladyna & Downing, 1989; Haladyna et al., 2002; Gronlund & Waugh, 2009; Nitko & Brookhart, 2011; Turgut & Baykul, 2012). It was found that only six of the articles included information about these rules and that these rules were taken into consideration, while in 26 articles only the items were written without any information. In addition, eight articles provided information that source books (e.g. textbooks of the Ministry of National Education, etc.) that can be considered valid in the subject areas of the articles were examined. Therefore, these eight articles were categorized as partial. When it was examined whether the opinions of field experts were taken regarding the appropriateness of the items, it was seen that expert opinions were not taken in only two of the articles. In addition to the specification tables, the analysis of techniques based on expert opinions can also be used to determine content validity. Eight of the articles employed the

analysis of different techniques based on expert opinions (Davis Technique (2), Kendal W Fit (1), Lawshe Technique (5)), while 32 of the articles did not utilize any statistical analysis. Furthermore, 14 of the articles took an expert opinion on face validity and reported the findings obtained, while 26 of the articles did not perform this process. The frequency values regarding the creation and implementation of the trial test form are given in Table 5.

**Table 5**

*Frequency Values Related to the Creation and Implementation of The Trial Test For*

| | **Yes** | | **Partial** | | **No/NI** | |
|---|---|---|---|---|---|---|
| | **f** | **%** | **f** | **%** | **f** | **%** |
| Information about the test instructions was given. | 9 | 22.5 | | | 31 | 77.5 |
| Information was given about the criteria considered in item ordering. | 1 | 97.5 | | | 39 | 2.5 |
| The formal features of the test form were mentioned. | 4 | 10 | 2 | 5 | 34 | 85 |
| The characteristics of the group to which the trial test form will be applied are described. | 19 | 47.5 | 11 | 27.5 | 10 | 25 |

After the writing stage of the items to be included in the trial test form, the stages of creating the form in which these items will be included and the realization of the pilot application follow. Information on this stage is given in Table 5. First of all, when it is analyzed whether information about the test instructions was given or not, it can be stated that only nine of the articles included this information, while 31 of them did not include any information. While information about the formal features during the preparation of the trial test form was included in four articles, no information was given in 34 articles. However, in two articles, the statement stating that “it was only formally organized” was used and no additional explanation was given. Therefore, these articles were evaluated as partial. For example, in article 39, the following expressions are stated: '... Creating the first draft form: The candidate achievement test, which was reorganized after the expert opinions, was formally organized and made ready for use by adding the instruction including the purpose and application method of the test." While the characteristics of the group to which the trial test form will be applied were included in 19 articles, they were not included in 10 articles. In 11 of the articles, it was seen that limited information about the group in which the application would be made was included under the title of study group and therefore it was evaluated as partial. Frequency values for item and test statistics are given in Table 6.

**Table 6**

*Frequency Values for Item and Test Statistics*

| | Yes | | Partial | | No/NI | |
|---|---|---|---|---|---|---|
| | **f** | **%** | **f** | **%** | **f** | **%** |
| Item difficulty index is included. | 39 | 97.5 | 1 | 2.5 | | |
| Average item difficulty is given. | 25 | 62.5 | | | 15 | 37.5 |
| Item discrimination index is given. | 39 | 97.5 | | | 1 | 2.5 |
| Average item discrimination is given. | 24 | 60 | | | 16 | 40 |
| Item standard deviation is given. | 9 | 22.5 | 1 | 2.5 | 30 | 75 |
| Item reliability index is given. | 8 | 20 | | | 32 | 80 |
| Option analysis was performed. | 2 | 5 | | | 38 | 95 |
| Information on item-total correlation was given. | 7 | 17.5 | | | 33 | 82.5 |
| The difference in the scores of the upper and lower groups was obtained by t-test. | 11 | 27.5 | | | 29 | 72.5 |
| The mean of the test is reported. | 23 | 57.5 | | | 17 | 42.5 |
| The standard deviation of the test is included. | 20 | 50 | | | 20 | 50 |
| The reliability of the test was reported. | 37 | 92.5 | | | 3 | 7.5 |
| A statistical package program was used to calculate item and test statistics. | 23 | 57.5 | | | 17 | 42.5 |
| Information about the items to be included in the test was given. | 25 | 62.5 | 9 | 22.5 | 6 | 15 |

The item and test statistics calculated from the data collected after the pilot application are highly valuable for guiding the selection of items to be included in the subsequent development of the achievement test. Information on the level of these statistics in the analyzed articles is given in Table 6. First of all, when item difficulty and discrimination indices are examined, it is seen that these indices are calculated and information is given in 39 out of 40 articles. While there is one article that does not include item discrimination indices, one article states that item difficulty indices are calculated but does not provide information about these indices, and is evaluated in the partial category. Another statistic that provides information about the quality of the items is the item standard deviation. While 30 of the articles did not report item standard deviation, nine of them included this value. The 30th-ranked article was evaluated as partial. Similar to the information on the item standard deviation, eight articles reported the item reliability index, while 32 articles did not provide any information. In multiple-choice items, the distribution of responses to options is very important in terms of the quality of the item. Therefore, performing option analysis in the process of developing an achievement test is one of the factors that increase the quality of the item and the test. Option analysis was conducted in only two of the articles, while it was not carried out in 38 articles. Item-total correlations, which provide information about the relationship between the scores obtained from the items in the test and the total score obtained from the entire test and provide information about whether the items sample similar behaviors

(Büyüköztürk, 2011), were calculated and reported in seven articles, while these values were not reported in 33 articles. Another indicator of the discrimination of the items is the difference between the item scores of the individuals in the lower and upper groups examined by independent samples t-test. While 11 of the articles calculated this statistic, 29 did not include it. The mean and standard deviation of the test also provide information about the scores obtained from the test and the group to which the test was applied. While 23 of the articles included the mean of the test, 17 did not include it, and half of the articles reported the standard deviation of the test. The number of articles reporting the findings on the reliability of the achievement tests is 37. There were three articles that did not share any values regarding the reliability of the test. 23 articles used various statistical package programs (TAP, SPSS, ITEMMAN) to calculate item and test statistics, while 17 did not report anything, so it was assumed that they were manually calculated. Finally, after the item and test statistics were performed, it was examined whether information about the items to be included in the test and their properties was provided. While 25 of the articles included this information, six did not provide any information. Nine articles in which the total number of items to be included in the test and the sequence numbers of the items were reported but the items were not specified or not shared in the appendix were evaluated in the partial category.

## Discussion and Results

In order to decide whether the new behaviors desired to be acquired through education have been learned at the expected level, these behaviors should be measured with measurement tools with sufficiently high validity and reliability (Kutlu & Altıntaş, 2021). One of the most frequently used measurement tools in education is achievement tests. In this study, a total of 40 articles that aim to develop achievement tests and whose titles include achievement test development were examined based on Turgut and Baykul (2012)'s achievement test development steps. The findings obtained in the study were discussed within the scope of the coding list developed by the researcher in line with expert opinions.

Having certain steps and systematics in the achievement test development process is important in terms of the quality of the test to be developed. In addition, the achievement test development process consists of planning, preparation, implementation, and reporting stages (Gömleksiz & Erkan, 2010). Therefore, it is important to have a plan for a qualified achievement test. It is thought that the absence of an achievement test development plan in

most of the articles analyzed may pose a problem both in terms of the achievement test developed and scientific reporting. The first step in test development is to determine the purpose for which the test scores will be used. Since decisions are made in line with the scores obtained from the tests and the decisions taken in education vary, it is necessary to determine the purpose in the process (Atılgan et al., 2015). In line with the findings, it can be said that the purpose of the test was clearly stated. However, the fact that the purpose of the test was confused with the purpose of the research in 11 articles, the purpose of the test was not specified, and it was not clear for what purpose the obtained scores would be used can lead to the result that the researchers did not understand this step.

Once the purpose of the test has been determined, the scope that will serve this purpose needs to be defined. The scope of the test refers to the learning products that will be tested by the items in the test. These learning products should be identified by counting them one by one and should be specified in the subjects or units in which the learning products are included. Only in this way can it be made clear what will be done while preparing the test and which scope the test will serve (Özçelik, 1997). In this context, it can be stated that in most of the articles examined, the learning products were clearly presented and the subjects and units belonging to them were specified. However, it is noteworthy that although there was a study on achievement test development, there were articles in which this step was not presented. The observation that all studies presented the number of items to be incorporated into the test, with nearly all (97.5%) specifying the item types, is considered as valuable input for achieving the intended goals of the developed tests. On the contrary, it can be asserted that the absence of test duration specification in the majority of studies (65%) and the lack of information on the number of options in some of them (35%) could potentially pose challenges for practitioners tasked with administering the test.

Another step in the achievement test development process is to reveal the distribution of learning products to the items, which gives content validity. Content validity can be defined as the sampling of the scope to be measured by test items (Baykul, 2000; Haynes et al., 1995; Mehrens & Lehmann, 1991). Therefore, content validity should not be ignored for a qualified and valid test (Demirel, 2006; Ebel, 1956; Lissitz & Samuelsen, 2007). For this step, preparing a table of specifications to ensure content validity is very important in achievement tests (Büyüköztürk et al., 2012). It can be concluded that this important step is provided in most of the articles analyzed (75%), but it is ignored in a considerable number (17.5%). The presence of articles (7.5%) that do not explicitly reference the specification table, which offers

crucial information for practitioners using the developed test, is considered a potential hindrance to the test's effectiveness in fulfilling its intended purpose. The submission of the prepared specification table to expert opinion provides information about whether or not the coverage is sampled before item writing and eliminates possible errors at this point (Gronlund, 1977). It was concluded that the majority of the studies (55%) skipped this step. However, the result that most of the articles (75%) wrote the items based on the specification table and almost all of the articles (95%) received expert opinion for the trial items is important in terms of producing a qualified achievement test. Moreover, the multiple-choice items used have certain writing rules that need to be taken into account. For these rules, particularly Haladyna et al., 2002 can be consulted. It is quite striking that in a very small portion of the articles (15%) these rules were taken into consideration, while in a large portion (65%) they were not mentioned at all. The finding that 20% of the articles were based on already written articles or reference books may lead to the conclusion that there may be more vulnerability to errors or incorrect item writing during article item. In addition to the creation of the specification table, various indices based on expert opinion can also be used to determine the content authenticity. The Lawshe (1975) technique, which quantifies the suitability of draft items to the relevant learning products with expert opinions, the Davis (1992) technique, which allows inferences to be made about candidate items by taking the opinions of at least three and at most 20 experts, or Kendall's coefficient of concordance are examples of such techniques. The finding that 20% of the articles utilized these indexes and included additional information may indicate that the use of these techniques is not widespread. The fact that each item is understood in the same way by practitioners and researchers is referred to as face validity (Mehrens & Lehmann, 1991; Nunnally & Bernstein, 1994). In line with the findings, it can be concluded that most of the articles (65%) ignored this type of validity.

It was observed that most of the articles did not include any information about the placement of the items in the trial test form, the features of this form and the piloting process before the pilot application was carried out after the trial item writing, and no information was shared. However, there are studies in the literature that discuss the effects of the location of the items in the test on the performance of individuals based on item response theory (Debeer & Janssen, 2013; Doğan Gül & Çokluk Bökeoğlu, 2018; Qian, 2014; Weirich et al., 2014; Weirich et al., 2017). Omitting this information can be viewed as a shortcoming in the respective articles.

Item analysis, which will be carried out with the pilot study data, is the process that provides information about whether the items measure the characteristics to be measured, and if not, what are the reasons for this and how they can be corrected (De Grutijter & Van der Kamp, 2008; DeVellis, 2006; Allen, 2012). Item analysis allows us to observe item characteristics and improve the quality of the test (Linn, 2011; Livingston, 2011). In order to develop a test consisting of quality items, item statistics are necessary. One of the item statistics calculated to determine the quality of the measurement tool is item difficulty. Item difficulty is defined as the ratio of the number of correct answers given to an item to the number of all respondents. Another item statistic is item discrimination. Item discrimination is the ability of an item to distinguish between individuals with high and low performance on the trait measured by the test (Linn, 2011; Livingston, 2011; Moses, 2017). These two statistics are the most frequently used indices in item analysis. The calculation and reporting of these indices in almost all of the articles (97.5%) can be considered as an important result in order to ensure the quality of the developed test. Item standard deviation and reliability indices are also statistics that provide information about item quality. It was concluded that these indices were not used as frequently as other indices in the articles analyzed. Especially after item difficulty and discrimination indices are determined, a very important component of item analysis is distractor analysis (Thissen et al., 1989). The main purpose of distractor analysis is to identify items in the measurement tool that need to be revised and renewed, to eliminate ineffective distractors and to increase the discriminatory power of multiple-choice items (Haladyna, 2016). With distractor analysis, it will be easier to identify possible errors in the items (Hingorjo & Jaleel, 2012). When the studies were examined, it was found that only 5% of the studies conducted distractor analysis, and it was concluded that this step, which is important for developing a quality achievement test, was skipped or not taken into consideration. Each item in an achievement test is expected to be related to other items. The correlation between each item score and the total score obtained from the whole test gives item-total correlations and these values reveal the relationship between the items and the whole test (Cristobal et al., 2007; Pallant, 2007). A large number of articles (82.5%) did not include these values and therefore, this lack of information about the tests makes it difficult to comment on the quality of the test. Another procedure that shows how well the items discriminate individuals in line with the trait to be measured is the comparison of the 27% lower and upper group averages for each item. It can be stated that the fact that this information was not included in most of the articles analyzed (72.5%) reveals that additional

evidence on discrimination was not presented. In addition to item analysis, the calculation of test statistics also provides important information about the characteristics of the test. Test statistics can also be calculated based on item statistics. Since items constitute the test, test statistics are a function of item statistics. One of the most frequently used of these statistics is the mean of the test and the other is the standard deviation of the test. The fact that half or nearly half of the articles do not include these values may indicate that test statistics may not be seen as a characteristic of the test. Another crucial concept in the test development process is test reliability, seen as an indicator of the extent to which measurement results are free from random errors. The fact that the reliability coefficient was calculated in almost all of the articles (92.5%) is a very important result in terms of revealing the quality of the developed test. While various methods for determining reliability exist in the literature (such as those involving multiple administrations, the test-halving method, etc.), it is noteworthy that the articles primarily relied on single-administration methods. In all but one article, the KR-20 reliability coefficient was calculated, with only one article using the halving test method. This pattern serves as a notable indicator of the prevalent methodological approach in use. For the calculation of item and test statistics, which are relatively difficult to calculate manually, there are different programs in the literature that require (EXCEL, R, etc.) or do not require (ITEMMAN, SPSS, TAP, etc.) formula writing. It was concluded that almost half of the articles (42.5%) used any program. Ultimately, it can be said that the purpose of developing an achievement test is to create a feasible measurement tool. In this context, sharing the items that emerged at the end of the achievement test development process and thus sharing the test is important in terms of serving this purpose. While the final version of the test is expected to be shared in the articles examined, the result that this sharing was realized in 62.5% of the articles can be considered as an indicator of the incomplete part of the studies.

In the articles reviewed as part of this research, it is evident that there are shortcomings at nearly every stage, beginning with the formulation of the test plan following the achievement test development process, all the way to the selection of items for inclusion in the final test form. It can be stated that there are similar problems in the studies examining the achievement test development processes carried out in the literature and that they are in parallel with this result (Boyraz, 2018; Karadağ, 2011; Mutluer & Yandı, 2012; Şahin et al., 2023).

## Recommendations

Based on the results obtained, especially when the aim of the research is to develop an achievement test, it can be characterized as a situation that should be meticulously completed in this process. In this context, it can be suggested that the achievement test development steps in the literature should be taken into consideration in the studies. Moreover, it is thought that the development of an achievement test development manual containing detailed explanations as a source of studies or revising the existing ones will also contribute to the literature. Although there are studies conducted to develop achievement tests in different fields, it is seen that more achievement tests have been developed in the fields of mathematics, science, and social studies, especially in science. For this reason, it is expected that conducting similar studies by limiting the field and expanding the years of research will contribute to the literature. Given that item difficulty and discrimination indices are derived from item analyses, and mean and standard deviation are computed from test analyses, while other item and test analyses are typically omitted, one of the recommendations of the study is to develop a freely accessible online platform that offers information and simplifies the calculation of these analyses.

## References

*Aksoy, Ş., & Özcan, H. (2020). Altıncı sınıf öğrencilerinin ses ve özellikleri ünitesi ile ilgili başarılarını ölçmeye yönelik bir test geliştirme çalışması. *Eğitimde Kuram ve Uygulama, 16*(2), 193-214.

*Aydın, E., & Selvi, M. (2020). Ortaokul öğrencilerine yönelik ekosistem, biyolojik çeşitlilik ve çevre sorunları başarı testinin geliştirilmesi. *Eğitim ve Toplum Araştırmaları Dergisi, 7*(2), 661-682.

*Aydın, T., & Faydaoğlu, Ş. (2022). Çember başarı testi geçerlik ve güvenirlik araştırması. *International Journal of New Trends in Arts, Sports & Science Education (IJTASE), 11*(4), 217-226.

*Bayırlı, H., & Köksal, H. (2022). 3. Sınıf hayat bilgisi dersi başarı testi geliştirme: geçerlik ve güvenirlik çalışması. *PESA Uluslararası Sosyal Araştırmalar Dergisi*, *8*(2), 86-99.

*Birgin, A., & Özcan, H. (2022). 8. Sınıf öğrencilerinin mevsimlerin oluşumu ile ilgili bilgilerini ölçmeye yönelik bir başarı testinin geliştirilmesi. *Trakya Üniversitesi Sosyal Bilimler Dergisi*, *24*(1), 305-326.

*Boz, S., Özcan, H., & Sarıoğlan, A. B. (2023). Ortaokul öğrencilerinin basınç konusu ile ilgili bilgilerini ölçmeye yönelik bir başarı testinin geliştirilmesi. *İnönü Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi, 10*(19), 14-29.

*Çiftcibaşı, F., Karamustafaoğlu, S., & Bolat, A. (2023). 'Güneş sistemi ve tutulmalar'ünitesine yönelik başarı testi geliştirilmesi. *Gaziantep Üniversitesi Eğitim Bilimleri Dergisi, 7*(1), 1-26.

*Dede, H., & Keleş, İ. H. (2020). Saf madde, karışımlar ve karışımların ayrılması konularında yaşam temelli başarı testinin geliştirilmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 40*(3), 797-825.

*Doğru, M., & Çepni, S. (2023). Karşılaştırmalı Olarak Geleneksel Çoktan Seçmeli ve Bağlam Temelli Başarı Testi Hazırlama Çalışması: 7. Sınıf Işığın Madde ile Etkileşimi Ünitesi. *Fen Matematik Girişimcilik ve Teknoloji Eğitimi Dergisi*, *6*(1), 74-101.

*Elbay, S. (2020). TC İnkılap Tarihi ve Atatürkçülük dersi 2. ünitesine yönelik kazanım odaklı başarı testi geliştirme çalışması. *E-Uluslararası Eğitim Araştırmaları Dergisi, 11*(1), 53-68.

*Emirtekin, E., Kışla, T., Polan, Ş., & Dönmez, O. (2020). Etkileşimli eğitsel video ve başarı testinin geliştirilmesi: IP adresi kavramı örneği. *Journal of Instructional Technologies and Teacher Education, 9*(1), 42-51.

*Eren, A. A., Önal, N. T., & Büyük, U. (2020). Elementler ve bileşikler konusu için geçerli ve güvenilir bir başarı testi geliştirme çalışması. *Pearson Journal, 5*(6), 152-167.

*Eroğlu, P., & Girgin, S. (2020). Ortaokul öğrencilerine yönelik biyoçeşitlilik başarı testi geliştirilmesi. *Uluslararası Sosyal ve Beşeri Bilimler Araştırma Dergisi, 7*(58), 2319-2326.

*Gül, A. C., Apaydın, Z., & Çobanoğlu, E. O. (2021). Canlılar dünyasına yolculuk konu alanına yönelik başarı testi geliştirme çalışması. *Ordu Üniversitesi Sosyal Bilimler Enstitüsü Sosyal Bilimler Araştırmaları Dergisi, 11*(1), 74-84.

*Güven, Ç., & Selvi, M. (2021). Beşinci sınıf "Elektrik Devre Elemanları" ünitesine yönelik başarı testi geliştirme. *International Journal of Current Approaches in Language, Education and Social Sciences, 3*(1).

*Hançer, M., Aydoğan, N., & Çankaya, Ö. (2021). Fen bilgisi öğretmen adaylarının temel laboratuvar fen bilgilerinin ölçülmesine yönelik başarı testi geliştirilmesi: geçerlik ve güvenirlik analizleri. *Uluslararası Eğitim Bilim ve Teknoloji Dergisi*, *7*(1), 57-76.

*Kahtalı, B. D., & Erdem, İ. (2020). Farklı tür metinler için dinlediğini anlama başarı testlerinin geliştirilmesi. *Electronic Turkish Studies, 15*(2).

*Kaplan, E., Bektas, O., & Karaca, M. (2022). Madde ve ısı ünitesi başarı testi geliştirme çalışması. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, (63), 78-116. DOI: 10.21764/maeuefd.985968

*Karakuyu, A., & Ocak, G. (2022). Dijital vatandaşlığa yönelik başarı testi geliştirme çalışması. *İnönü Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi, 9*(18), 32-42.

*Kargın, P. D., & Gül, Ş. (2021). Altıncı sınıf "Vücudumuzdaki Sistemler ve Sağlığı" ünitesine yönelik bir başarı testi geliştirilmesi. *Ihlara Eğitim Araştırmaları Dergisi, 6*(1), 1-26.

*Kaya, S., & Gül, Ş. (2020). 11. Sınıflar için 'Sindirim Sistemi' konusuna yönelik başarı testi geliştirme çalışması. *Uluslararası Sosyal ve Eğitim Bilimleri Dergisi, 7*(13), 72-97.

*Kılıç, Ç., & Girgin, S. (2022). 7. Sınıf hücre ve bölünmeler ünitesi akademik başarı testi geliştirilmesi: geçerlik ve güvenirlik analizi. *International Journal Of Socıal Humanıtıes Scıences Research, 9*(81), 407-420.

*Kurt, A., Aydın, M., & Bekereci, Ü. (2023). 6. Sınıf ses ve özellikleri ünitesine yönelik başarı testi geliştirme çalışması: geçerlik ve güvenirlik analizi. *E-International Journal of Educational Research, 14*(3).

*Laçin, E. (2022). Okul öncesi öğretmenlerinin erken okuryazarlık bilgi düzeyini ölçmeye yönelik bilgi testi geliştirme çalışması. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, (53), 150-172.

*Meriçelli, M., & Güyer, T. (2020). Enformatik dersi için başarı testi geliştirme çalışması: Güvenilirlik ve geçerlilik işlemleri. *Kastamonu Eğitim Dergisi, 28*(1), 549-557.

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

*Özcan, H., Boz, C., & Özkaya, A. (2020). 7. Sınıf öğrencilerinin hücre konusuyla ilgili anlayışlarını ölçmeye yönelik bir test geliştirme çalışması. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 17*(46), 203-233.

*Özcan, H., Çetinkaya, İ., & Arık, S. (2021). Ortaokul öğrencilerinin basit makineler ünitesi ile ilgili anlayışlarını ölçmeye yönelik bir test geliştirme çalışması. *Muğla Sıtkı Koçman Üniversitesi Eğitim Fakültesi Dergisi, 8*(1), 16-36.

*Özkan, S., & Yadigaroğlu, M. (2020). Başarı testi geliştirme: asit-baz başarı testi geçerlik ve güvenirlik araştırması, *Turkish Studies-Education, 15*(2), 1141-1163. https://dx.doi.org/10.29228/TurkishStudies.41734

*Özkılıç, G. E., Bektas, O., & Karaca, M. (2023). Sindirim Sistemi Ünitesine Yönelik Başarı Testi Geliştirme: Geçerlik ve Güvenirlik Çalışması. *Araştırma ve Deneyim Dergisi, 8*(1), 115-154.

*Pazar, B., & Karamustafaoğlu, S. (2023). "Saf Madde ve Karışımlar" Ünitesi başarı testi geliştirme: Geçerlik ve güvenirlik. *Anadolu University Journal of Education Faculty, 7*(2), 404-432.

*Sevim, S., Uysal, İ., & Demirci, E. (2021). Fen bilimleri dersi 5. sınıf "Işığın Yayılması" ünitesine yönelik başarı testi geliştirme çalışması. *Caucasian Journal of Science, 8*(2), 224-246.

*Sontay, G., & Karamustafaoğlu, O. (2020). Fen bilimleri dersi "Güneş, Dünya ve Ay" ünitesine yönelik başarı testinin geliştirilmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 40*(2), 511-551.

*Süslü, A., & Ötken, Ş. (2020). Uluslararası yolcu taşımacılığı (SRC1) sürücü mesleki yeterlilik sınavlarına ait bir başarı testi geliştirilmesi: Geçerlik ve güvenirlik çalışması. *Stratejik ve Sosyal Araştırmalar Dergisi, 4*(1), 135-145.

*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.

*Topay, N., & Yılmaz, M. (2023). Biyoloji ve fen bilgisi öğretmenlerine yönelik tamamlayıcı ölçme ve değerlendirme teknikleri başarı testi geliştirilmesi. *Gazi Eğitim Bilimleri Dergisi, 9*(2), 214-240.

*Uçar, R., & Aktamış, H. (2019). Astronomi'ye yönelik tutum ölçeği ve 7. sınıf "Güneş Sistemi ve Ötesi" ünitesine yönelik başarı testi geliştirme çalışması, *Batı Anadolu Eğitim Bilimleri Dergisi, 10*(1), 57-79.

*Üçüncü, G., & Sakiz, G. (2020). Başarı testi geliştirme süreci: İlkokul dördüncü sınıf maddeyi tanıyalım ünitesi örneği. *Kastamonu Eğitim Dergisi, 28*(1), 82-94.

*Yılmaz, İ., & Yılmaz, E. (2021). 4. Sınıf matematik dersi doğal sayılar alt öğrenme alanına ilişkin başarı testi geliştirme çalışması. *Kesit Akademi Dergisi, 7*(26), 295-310.

*Zeyrek, A. Ş., Kurban, N. K., & Arslan, S. (2020). Bir başarı testi geliştirme çalışması: hemşirelik öğrencilerinin intramüsküler enjeksiyon becerilerini ölçme. *Gümüşhane Üniversitesi Sağlık Bilimleri Dergisi*, *9*(2), 133-141.

Açıkgöz, M., & Karslı, F. (2015). Alternatif Ölçme-Değerlendirme Yaklaşımları Kullanılarak İş ve Enerji Konusunda Geliştirilen Başarı Testinin Geçerlilik ve Güvenirlik Analizi. *Amasya Üniversitesi Eğitim Fakültesi Dergisi*, *4*(1), 1-25.

Airasian, P. W. (2001). *Classroom assessment: Concepts and applications*. McGraw-Hill.

Akbulut, H. İ., & Çepni, S. (2013). Bir üniteye yönelik başarı testi nasıl geliştirilir?: İlköğretim 7. sınıf kuvvet ve hareket ünitesine yönelik bir çalışma. *Amasya Üniversitesi Eğitim Fakültesi Dergisi, 2*(1), 18-44.

Akkuş, R., & Akkaş, E. N. (2021). Ortaokul 5., 6. ve 7. sınıf seviyelerinde matematik genel başarı testleri geliştirme çalışması. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, (51), 180-209.

Allen, D. D. (2012). Validity and Reliability of the Movement Ability Measue: A Self-Report Instrument Proposed for Assessing Movement Across Diagnoses and Ability Levels. *Physical Thearapy, 87*(7), 899-916

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational objectives*. Longman.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of educational research, 42*(2), 145-170. https://doi.org/10.3102/00346543042002145.

Anderson, L.W. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. Lawrence Erlbaum Associates, Inc.

Armağan, F. Ö., & Demir, N. (2019). Astronomi başarı testi geliştirme: Geçerlik ve güvenirlik çalışması. *Maarif Mektepleri Uluslararası Eğitim Bilimleri Dergisi*, *3*(1), 52-70.

Atılgan, H., Kan, A., ve Doğan, N. (2015). *Eğitimde ölçme ve değerlendirme*. Hakan Atılgan (Ed.), Test geliştirme (316-348). Anı Yayıncılık.

Aydin, F. (2018). L2 metalinguistic knowledge and L2 achievement among intermediate-level adult Turkish EFL learners. *Journal of Language and Linguistic Studies, 14*(1), 28-49.

Ayvacı, H. Ş., & Durmuş, A. (2016). Bir başarı testi geliştirme çalışması: Isı ve sıcaklık başarı testi geçerlik ve güvenirlik araştırması. *Ondokuz Mayis University Journal of Education Faculty, 35*(1), 87-103.

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

Bolat, A., & Karamustafaoğlu, S. (2019). "Vücudumuzdaki Sistemler" Ünitesi başarı testi geliştirme: geçerlik ve güvenirlik. *Gazi Eğitim Bilimleri Dergisi*, *5*(2), 131-159.

Boyraz, C. (2018). Investigation of achievement tests used in doctoral dissertations department of primary education (2012-2017). *Inonu University Journal of the Faculty of Education, 19*(3), 14-28. doi: 10.17679/inuefd.327321.

Brookhart, S. M., & Nitko, A. J. (2011). Strategies for Constructing Assessments of Higher-Order Thinking Skills. *Assessment of Higher Order Thinking Skills, 1*, 327-59.

Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students*. Pearson.

Büyüköztürk, Ş. (2011). *Sosyal bilimler için veri analizi el kitabı*. Pegam Akademi Yayıncılık.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2012). *Bilimsel araştırma yöntemleri (13. Baskı).* Pegem Akademi.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Allyn and Bacon.

Cheung, D. (2003). Guidelines for writing multiple-choice items. *Hong Kong Science Teachers' Journal, 21*(2), 1-11.

Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment*. McGrawHill Higher Education.

Cristobal, E., Flavian, C., & Guinaliu, M., (2007). Perceived e-service quality (PeSQ): measurement validation and effects on consumer satisfiction and web site loyalty, *Journal of Service Theory and Practice, 17*(3), 317-340

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

Davis L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*, 194-197.

De Grutijter, D. N., & Van der Kamp, L. J. (2008). Statistical Test Theory for the Behavioral Sciences. Chapman&Hall, London, 280p.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185.

Demirel, Ö. (2006). *Öğretimde planlama ve değerlendirme öğretme sanatı*. Pegem Yayıncılık.

DeVellis, R. F. (2006). Quantitative and ıssues and approaches: Classical Test Theory (CTT) and Item Response Theory (IRT). *Medical Care, 44*(11), 50-59.

Doğan Gül, Ç., & Çokluk Bökeoğlu, Ö. (2018). The comparison of academic success of students with low and high anxiety levels in tests varying in item difficulty. *Inonu University Journal of the Faculty of Education, 19*(3), 252-265. https://doi.org/10.17679 /inuefd.341477

Ebel, R. L. (1956). *Essentials of educational measurement (1st Ed.).* Prentice-Hall.

Ersoy, E., & Bayraktar, G. (2018). İlkokul 4. sınıf matematik dersi "Ondalık Gösterim" alt öğrenme alanına ilişkin başarı testi geliştirilmesi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, (46), 240-266.

Fellenz, M. R. (2004). Using assessment to support higher level learning: the multiple choice item development assignment. *Assessment & Evaluation in Higher Education, 29*(6), 703-719.

Gömleksiz, M., & Erkan, S. (2010). *Eğitimde ölçme ve değerlendirme* (2. Baskı). Nobel Yayın Dağıtım.

Gronlund, N. E. (1977). *Constructing achievement test*. Prentice Hall, Inc.

Gronlund, N.E., & Waugh, C.K. (2009). *Assessment of Student Achievement*. Pearson Education.

Güven, E. (2013). Çevre sorunları başarı testinin geliştirilmesi ve öğretmen adaylarının bilgi düzeylerinin belirlenmesi. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, *3*(2).

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Allyn and Bacon

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, *2*(1), 37-50.

Haladyna, T. M., (2016). *Item analysis for selected response test ıtems*. Handbook of Test Development. Lane, S., Raymond, M. and Haladyna, T. (Eds.), Routledge, 392-409.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333.

Haladyna, T.M. (2004). *Developing and validating multiplechoice test items,* Lawrence Erlbaum Associates, Publishers.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment*, *7*(3), 238.

Heckman J.J., & Kautz T. (2012). Hard evidence on soft skills. *Labour Econ 19*(4), 451–464.5.

Heckman J.J., & Kautz T. (2014). Fostering and measuring skills: Interventions that improvecharacter and cognition. In Heckman JJ, Humphries JE, & Kautz T. (Eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life*, 341–430. Univ of Chicago Press.

Hingorjo, M. R., & Jaleel, F., 2012. Analysis of one best MCQs: The difficulty, discrimination ındex and discrimination efficiency, *Journal of Pakistan Medical Association*, *62*(2), 142-147.

Irwing P., & Hughes D. J. (2018). Test development. In Irwing P., Booth T., & Hughes D. J. (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, 3–48. Hoboken, Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch1.

Karaboğaz, Y., & Ergene, Ö. (2023). Beceri Temelli Orantısal Akıl Yürütme Başarı Testinin Geliştirilmesi. *Journal of Individual Differences in Education*, *5*(1), 31-47.

Karadağ, E. (2011). Eğitim bilimleri doktora tezlerinde kullanılan ölçme araçları: Nitelik düzeyleri ve analitik hata tipleri. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 311-334.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy*, *65*(23), 2276-2284.

Kline, P. (2000). *Handbook of psychological testing*. Routledge.

Koç, N. (1985). Standart başarı testlerinin, bir eğitim sisteminde verilen çeşitli kararlardaki yeri ve önemi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi. 17*(01),159-172. https://doi.org/10.1501/Egifak_0000001047

Kubiszyn, T., & Borich, G. (1996). *Educational testing and measurement: Classroom application and practice*. Harper Collins.

Kutlu, Ö., & Altıntaş, Ö. (2021). Psikolojik ölçmelerin kısa tarihi ve 21. Yüzyılda sınıf içi durum belirleme anlayışı. *Trakya Eğitim Dergisi*, *11*(3), 1599-1620.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). *Handbook of test development*. Routledge.

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

Lehmann, I. J., & Mehrens, W. A. (1991). *Measurement and evaluation in education and psychology*. *Holt*.

Linn, R. L. (2011). *Handbook of test development*. Routledge.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational researcher*, *36*(8), 437-448.

Mayring, P. (2004). Qualitative content analysis. *A companion to qualitative research*, *1*(2), 159-176.

Mehrens, W. A. & Lehmann, I.J. (1991), *Measurement and Evaluation in Education and Psychology*. Harcourt Brace Jonanovich: Forth Worth.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Miller, M., Linn, R., & Gronlund, N. (2012). *Measurement and assessment in teaching (11th Ed.)*. Pearson

Moses, T. (2017). A review of developments and applications in item analysis. *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS*, 19-46.

Mutluer, C., & Yandı, A. (2012, September). Türkiye'deki üniversitelerde 2010-2012 yılları arasında yayımlanan tezlerdeki başarı testlerin incelenmesi. Paper presented at the Eğitimde ve Psikolojide Ölçme ve Değerlendirme III. Ulusal Kongresi, Turkey: Bolu. Abstract retrieved from https://www.epodder.org/wpcontent/uploads/2020/07/EPOD-2012.pdf.

Narlı, S., & Başer, N. (2008). "Küme, Bağıntı, Fonksiyon" konularında bir başarı testi geliştirme ve bu başarı testi ile üniversite matematik bölümü 1. sınıf öğrencileirnin bu konulardak hazırbulunuşluklarını betimleme üzerine nicel bir araştırma. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi, 24*, 147-158.

Nunnally J.C., & Bernstein I.H. (1994). *Psychometric theory*. McGrawHill.

Özçelik, D. A. (1997). *Test Hazırlama Klavuzu (3. Baskı).* ÖSYM Eğitim Yayınları.

Özçelik, D. A. (2010). *Ölçme ve değerlendirme*. Pegem

Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows (3rd ed.)*. Open University Press.

Popham, W. J. (2008). *Transformative assessment*. ASCD.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational researcher*, *31*(7), 3-14.

Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement, 38*(7), 518-534. https://doi.org/10.1177/014662161453431 2

Saadat, S., Noori, M., Alipour-Anbarani, M., Mousavi Bazaz, N., Babakhanian, M., Montazeri KHadem, A., & Azadi, H. (2021). Students' Challenge in Answer-changing on Multiple-choice Exams; Doubting the Answer or Not? A Systematic Review. *Medical Education Bulletin*, *2*(1), 137-144.

Saraç, H. (2018). Fen bilimleri dersi 'maddenin değişimi'ünitesi ile ilgili başarı testi geliştirme: Geçerlik ve güvenirlik çalışması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 18*(1), 416-445.

Singh, C., & Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts. *American Journal of Physics, 71*(6), 607-617.

Şahin, M. G., Yıldırım, Y., & Öztürk, N. B. (2023). Examining the Achievement Test Development Process in the Educational Studies. *Participatory Educational Research*, *10*(1), 251-274.

Şahin, M., Başkurt, İ., & Deringöl, Y. (2023). İlkokul 3. Sınıf Öğrencilerine Yönelik Matematik Problem Çözme Başarı Testinin Geliştirilmesi. *Bayburt Eğitim Fakültesi Dergisi*, *18*(39), 811-838.

Şen, H. C., & Eryılmaz, A. (2011). Bir başari testi geliştirme çalişmasi: Basit elektrik devreleri başari testi geçerlik ve güvenirlik araştirmasi. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *8*(1), 1-39.

Tavşancıl, E., & Aslan, E. (2001). *Sözel, yazılı ve diğer materyaller için içerik analizi ve uygulama örnekleri*. Epsilon Yayınevi.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247- 260.

Thorndike, R. M., & Thorndike-Christ, T. M. (2013). *Measurement and evaluation in psychology and education*. Pearson Higher Ed.

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme (4. Baskı)*. Pegem Akademi.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6.* National İnstitute Science Education.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*, 535-548. https://doi.org/ 10.1177/0146621614534955.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115-129. https://doi.org/10.1177/0146621616676791.

Yeşilyurt, E. (2012). Öğretmen adaylarının bilişsel alanla ilgili sınama durumu soruları yazma yeterliklerinin değerlendirilmesi. *Kastamonu Eğitim Dergisi, 20*(2), 519-53.

Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri. (9. Baskı).* SeçkinYayıncılık.

Yin, R. K. (2014). *Case study research: Design and methods*. Sage Publication.