Id: 1169497

Participatory Educational Research (PER)
Vol.10(1), pp. 251-274, January 2023
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.23.14.10.1

# Examining the Achievement Test Development Process in the Educational Studies

Melek Gülşah Şahin
*Assessment and Evaluation in Education, Gazi University, Ankara, Türkiye*
*ORCID: 0000-0001-5139-9777*

Yıldız Yıldırım [*]
*Measurement and Evaluation in Education, Aydin Adnan Menderes University, Aydin, Türkiye*
*ORCID: 0000-0001-8434-5062*

Nagihan Boztunç Öztürk
*Lifelong Learning Center, Hacettepe University, Ankara, Türkiye*
*ORCID: 0000-0002-2777-5311*

**Article history**
**Received:**
01.08.2022

**Received in revised form:**
30.10.2022

**Accepted:**
18.12.2022

**Keywords:**
Achievement test; mathematics; assessment and evaluation; achievement test development; document analysis

Literature review shows that the development process of an achievement test is mainly investigated in dissertations. Moreover, preparing a form that will shed light on developing an achievement test is expected to guide those who will administer the test. In this line, the current study aims to create an "Achievement Test Development Process Control Form" and investigate the achievement tests for Maths based on this form. Document analysis was conducted within the framework of qualitative research and was done based on descriptive analysis. Within the scope of the research, 1683 articles published in designated journals between 2015-2020 were reviewed. It was determined that a mathematics achievement test was developed in 39 of these articles, which were coded on the control form. The articles that were included in the scope of the current study were investigated in terms of the type of items used in the tests, the theory or practice on which the test was developed, the use of rubric for open-ended items, the number of items in the pilot and final form, features of the test form as well as those pertaining to the table of specifications, the features of item pool, the evaluation of pilot testing, the evaluation of real study, test validity and reliability, and the setting in which tests were administered. The current study findings show that mostly an item pool was not prepared, the pilot application was not conducted or was not specified, and even if it was conducted, item analysis was not performed, test forms or example items were not included in the articles, and there were some deficiencies regarding validity. On the other hand, it was clear that the articles mostly specified the test goal and reported the reliability coefficient. In light of the current

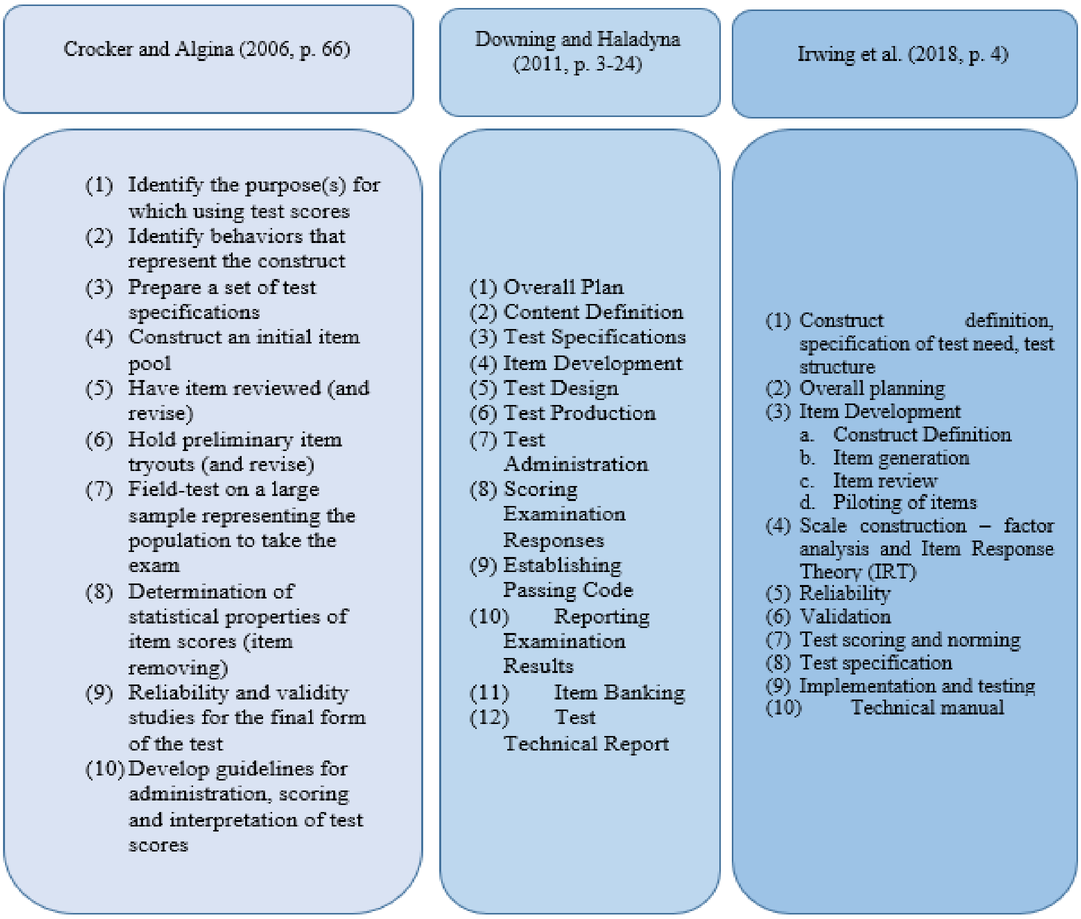
[*] Correspondency: yildizyldrm@gmail.com

findings, some suggestions are provided for test developers and those who will administer these tests.

## Introduction

Educational assessment and evaluation can be used for the purposes of diagnosis and formalization as well as identifying one's level, while it has been one of the leading ways to identify students' level of achievement. The concept of achievement is related to the changes in cognitive behaviours that can be altered through teaching or education, and it is specifically associated with three fundamental concepts: Knowledge, skills, and abilities. The concept of knowledge requires remembering or understanding for learning principles and facts. The concept of skill refers to cases that can be observed and includes a performance. Lastly, the concept of ability means the use of knowledge and skills together, while its development requires a long time period (Haladyna, 2004). Considering the system approach in education, it is possible to identify students' achievement for the behaviours determined in accordance with the elements of input, process, and outcome. For example, identifying students' level of readiness serves the goal of evaluating the input of educational system. On the other hand, questioning the level of achieving learning at the end of a class serves the goal of identifying the element of output. As constructive approach is commonly adopted today, the concept of formative evaluation has turned out to be more important, and the main goal is now to monitor students' learning. Identifying students' level of achievement is also helpful about monitoring students' learning throughout the process.

It is obvious that achievement tests are commonly used in order to identify students' success. Tests are assessment and evaluation instruments used for identifying the amount or degree of learning numerically in an environment designed by developers (Haladyna, 2004). Cronbach (1990, p. 37) divides tests into two as typical response tests and maximum performance tests. Typical response tests measure psychological constructs such as attitude, perception, personality, motivation, or interest, while such tests do not include items with a correct answer. In such tests, individuals generally focus on self-reporting rather than answering the item correctly. On the other hand, maximum performance tests include achievement tests, intelligence tests and aptitude tests. In these tests, individuals are expected to display the highest performance. Considered as one of the maximum performance tests, achievement tests can be developed by teachers, or they can be standard tests. It is of vital importance to follow the steps of developing an achievement test in order to ensure that achievement tests measure the intended type of success in line with its purpose and to be sure that they are free of mistakes as much as possible. While there are quite many resources that define the process of developing an achievement test in detail, Figure 1 shows the processes suggested by Crocker and Algina (2006, p. 66); Downing and Haladyna (2011, p. 3-24) and Irwing et al. (2018, p. 4).



| Crocker and Algina (2006, p. 66) | Downing and Haladyna (2011, p. 3-24) | Irwing et al. (2018, p. 4) |
|---|---|---|
| (1) Identify the purpose(s) for which using test scores<br>(2) Identify behaviors that represent the construct<br>(3) Prepare a set of test specifications<br>(4) Construct an initial item pool<br>(5) Have item reviewed (and revise)<br>(6) Hold preliminary item tryouts (and revise)<br>(7) Field-test on a large sample representing the population to take the exam<br>(8) Determination of statistical properties of item scores (item removing)<br>(9) Reliability and validity studies for the final form of the test<br>(10) Develop guidelines for administration, scoring and interpretation of test scores | (1) Overall Plan<br>(2) Content Definition<br>(3) Test Specifications<br>(4) Item Development<br>(5) Test Design<br>(6) Test Production<br>(7) Test Administration<br>(8) Scoring Examination Responses<br>(9) Establishing Passing Code<br>(10) Reporting Examination Results<br>(11) Item Banking<br>(12) Test Technical Report | (1) Construct definition, specification of test need, test structure<br>(2) Overall planning<br>(3) Item Development<br>a. Construct Definition<br>b. Item generation<br>c. Item review<br>d. Piloting of items<br>(4) Scale construction – factor analysis and Item Response Theory (IRT)<br>(5) Reliability<br>(6) Validation<br>(7) Test scoring and norming<br>(8) Test specification<br>(9) Implementation and testing<br>(10) Technical manual |

Figure 1. Achievement test development process.

Considering the processes of developing an achievement test given in Figure 1, it is obvious that all resources depict a similar flow. Furthermore, there are some standards suggested by widely-acknowledged bodies about developing an achievement test such as American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME). Figure 2 shows the process of developing an achievement test in line with the given standards.

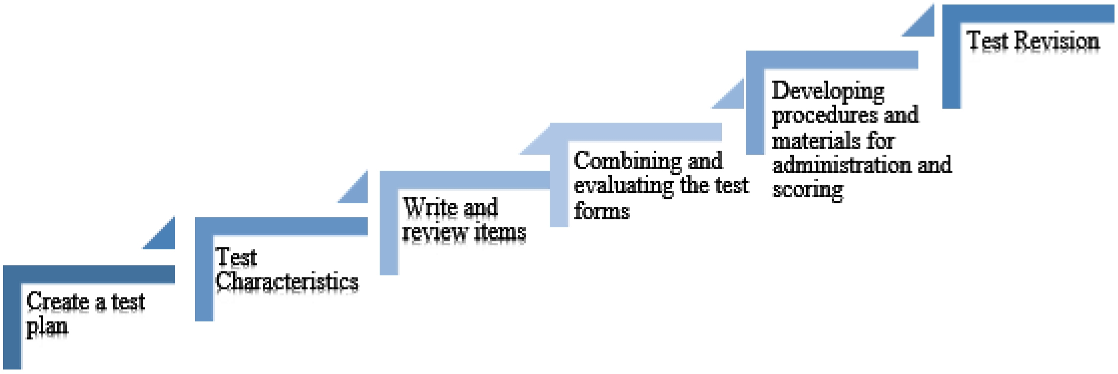

Create a test plan
Test Characteristics
Write and review items
Combining and evaluating the test forms
Developing procedures and materials for administration and scoring
Test Revision

Figure 2. The steps of the process of developing a standard achievement test.

According to Figure 2, the process of developing a test requires a systematic approach regarding the validity of the test scores and the decisions based on this, which makes it a priority to create a test plan (Lane et al., 2016). In other words, creating a test plan and following the steps included in the plan provide test developers with a standard as well as evidence of validity. Standards for Educational and Psychological Testing list five steps to be followed while developing a test (AERA et al., 2014). The first of these steps is to determine the test characteristics. At this step, the issues to decide on include goal of the test, content of the test, item format (question, performance task, etc.), how to receive responses, how to conduct scoring, length of the test, procedure to administer the test (paper-based, computer-based), systems to develop the items and test. The goal of a test can be to choose, reveal proficiency, classify, and do screening (Turgut, 1992). The scope of the exam, its degree of difficulty, the format of the items to be included in the exam, the length of the exam, and alike can all change in accordance with the goal of the test.

The second step of developing a test is to write and review items. At this stage, it is necessary to create an item pool, review items and conduct a pilot testing, do the analysis regarding the pilot testing, create a rubric if needed due to the item format and examine scorer reliability. Creating a table of specifications at this stage is of vital importance for content validity. Table of specifications can be considered as a more specified version of the test plan because it is a draft plan about what to do until the test form is created (Çetin, 2019). Table of specifications includes topics, learning outcomes, qualities that students will acquire, taxonomic level and number of questions per learning outcome. Another important decision in pilot testing is how to determine the sample. The first thing to do here is to choose a representative sample. While it is appropriate to have a randomly chosen sample group with enough participants as the students constitute a homogeneous structure, stratified sampling method can be adopted if the group of students is heterogenous or if it is possible to choose students from schools of different success levels. There are different criteria in the literature to determine the number of students. According to these different criteria, the number of students can vary between 120 and 400 (Özçelik, 1992), it is necessary to have a large group of sample with 100 to 200 participants (Crocker & Algina, 2006) and it is important to proceed with care when the sample is larger than 200 participants (Haladyna, 2004).

At the end of the pilot testing, it is necessary to investigate the validity and reliability of the test. At this stage, it is possible to obtain item and test statistics. The statistics to be calculated at this stage include item discrimination index, item difficulty index, distractor analysis, test reliability and scorer reliability.

Combining and evaluating the test forms is the third step of developing a test. It is important to ensure that the items are located on each test form in line with the appropriate rules (not including clues or not including similar items on the same test form, and so on) after receiving expert opinions. The fourth step requires developing procedures and materials for administration and scoring. At this stage, an instruction to administer the test is prepared, and some reliability procedures are developed both for administering and scoring the test. The last step is test revision. At this stage, the test is revised through reviewing test instructions on how to administer the test and answer the items in the test periodically, making the necessary changes if there is a change of content in the test, reflecting the changes about the curriculum on the test, changing the time to administer the test if necessary.

As is clear, developing a test consists of a series of steps that should be followed carefully. These steps about developing, administering, and evaluating a test should be followed meticulously for the reliability and validity of the decisions made about the students. First and foremost, test developers need to identify the goal of the test as well as the qualities of the group for whom the test will be used. At this stage, there are some other significant elements to be considered such as experiences of the test developers, receiving expert opinion from different fields and examining the experiences of the group for whom the test will be used. Shortly, collecting evidence for the goal is an important component of test development. Depending on the recently combined concept of validity about collecting the necessary evidence to identify validity (Messick, 1995), researchers can consider combining the evidence as to the meaning, interpretation and use of the test scores. Unlike traditional validity, combined validity does not differentiate content, criterion, and construct validity. Considering the combined validity based on collecting evidence, collecting evidence only for content validity, which is a kind of traditional validity, in the process of developing an achievement test can be a mistaken and deficient approach. In this line, during the process of developing an achievement test, it is important to collect evidence on what test scores mean, how they will be interpreted and used. The evidence can include preparing a table of specifications, receiving expert opinions, etc. as well as confirmatory factor analysis (CFA), explanatory factor analysis (EFA), examining correlation with a converging or discriminating structure, and others.

Literature review shows that different measurement tools are developed/adapted and administered for different purposes in different fields. Most frequently preferred tools are achievement tests and scales. There are also quite many studies in the literature that focus on examining the qualities of the processes of developing and adapting a scale (Acar-Güvenir & Özer-Özkan, 2015; Boztunç-Öztürk et al., 2015; Delice & Ergene, 2015; Ergene, 2020; Mor-Dirlik, 2014; Şahin & Boztunç-Öztürk, 2018). However, literature review shows that there is a limited number of studies that focus on examining test development in terms of the necessary criteria. Evrekli et al. (2011) conducted a study to examine graduate dissertations in the field of science education for the aim of identifying the related deficiencies. They developed a form called “Thesis Evaluation Form” to examine the method part of the dissertations. Also, Karadağ (2011) carried out a study to examine the qualities of assessment tools used in doctoral thesis in the field of educational sciences and the type of mistakes they had. The researcher developed a form called “Evaluation Form” to examine 5 items in the dimension of achievement test in order to identify the general quality of the measurement tools as well as their qualities differing according to the type. Mutluer and Yandı (2012) aimed at examining the steps of developing achievement tests included in graduate dissertations. Boyraz (2018) also carried out a study to examine the steps of developing an achievement test in doctoral dissertations in the field of primary education. In this line, the researcher developed an 8-category “Achievement Test Development Rubric” and examined the tests of multiple-choice questions according to this rubric.

Literature review shows that studies on developing an achievement test mostly focus on describing the process of developing an achievement test in dissertations. Unlike other studies in the literature, the current study aims at describing the process of developing an achievement test in an article. Examining the articles included in the current study will reveal the facts about the published studies, which will contribute to the literature in terms of identifying the deficiencies or limitations of the process of developing a test. Furthermore, the “Achievement Test Development Processes Control Form (ATDPCF)” created to examine the articles will contribute to teachers and researchers who develop a test in their field in terms of providing guidance.

## Method

### *Type of the Study*

The current study aims at developing an achievement test assessment form and examine the articles that focused on developing achievement tests for maths. In this line, the researchers reviewed the literature on the process of developing an achievement test and created a preliminary form. After that, expert opinion was taken from five experts for this form in the context of content validity. Revised in line with the expert opinions, the form was updated based on the studies in the literature developing a maths achievement test after examining articles published in some Turkish journals indexed in Social Sciences Citation Index (SSCI) and Emerging Sources Citation Index (ESCI). Furthermore, this form revealed the recent trends in developing a maths achievement test appearing in the articles included in the indexed journals. As the form was revised after reviewing the literature, the current study is a study of document analysis, which is a qualitative study. Additionally, it is a descriptive study as the researchers reviewed the articles that developed a maths achievement test and coded them according to the created form for the purpose of revealing the process of developing an achievement test.

### *Study Group*

In line with the purpose of the first stage of the current study, the researchers reviewed 1683 articles published in Education and Science Journal (Journal A), International Journal of Assessment Tools in Education (Journal B), Eurasian Journal of Educational Research (Journal C), Pegem Journal of Education and Instruction (Journal D) and Hacettepe University Journal of Education (Journal E) between the years 2015-2020. It was determined that the area where most tests were developed was mathematics (f=39; 31,45%) from these articles. For this reason, we choose the articles for which the mathematics test was developed as the study group. The researchers firstly reviewed the full text version of 1683 articles in order to choose the final data, chose 39 articles (Appendices 1) that developed an achievement test for maths and coded them according to ATDPCF. Table 1 shows the frequency and percentage distribution for the articles published in the indexed journals.

Table 1. The frequency and percentage distribution table for the articles published in the indexed journals

| **Journal Code** | **Total number of article (%)** | **The number of articles that developed an achievement test in maths (%)** | **The ratio of articles that developed maths achievement test to total articles** |
|---|---|---|---|
| Journal A | 510(%30,30) | 7(%17,95) | %1,37 |
| Journal B | 181(%10,75) | 1(%2,56) | %0,55 |
| Journal C | 414(%24,60) | 9(%23,08) | %2,17 |
| Journal D | 203(%12,06) | 10(%25,64) | %4,93 |
| Journal E | 375(%22,28) | 12(%30,77) | %3,20 |
| Total | 1683 | 39 | |

As is seen in Table 1, when the number of articles developing a maths achievement test was compared to the number of all articles, it was clear that Journal D had the highest proportion with 4,93% and Journal B had the lowest proportion with 0,55%.

### *Data Collection Tools*

The statements included in ATDPCF were created by the researchers after reviewing the literature. The first version of the form was reviewed by five experts of educational assessment and evaluation, and the form was finalized to be ready for use to examine the articles after making the necessary changes according to the expert opinions. Moreover, new statements were added to the form after receiving expert opinions when some criteria that were not included in the form appeared during the review. The final version of the form according to which coding was conducted included four parts. The first part included a record of the article (title of article, writer(s) etc.), the second part included some demographic information about the journal (the number of articles that developed an achievement test in maths / in all fields etc.), the third part included some general information on an achievement test (type of the item in the test, number of items in the pilot and final form etc.) and the last part included some details on the process of developing an achievement test (creating item pool, evidence of validity and reliability and so forth).

### *Data Collection Techniques*

The data were collected via reviewing articles published in Journal A, B, C, D and E between the years 2015-2020. The reason why the researchers chose the aforementioned journals is that they are journals published in Turkey and indexed in Web of Science (WOS), SSCI or ESCI. WOS was the preferred index because it is widely recognized in the field of education all around the world at an international level. During this process, the articles published in the mentioned journals that developed an achievement test in maths and in other fields were filed separately. 39 of the 1683 articles that were reviewed developed an achievement test for maths. Moreover, the number of articles that developed an achievement test in other fields is 85. Articles were coded by the two coders independently in order to calculate percentage of agreement (Miles & Huberman, 1994, p. 64). The percentage of agreement calculated to be %99,05, which means that inter-coder reliability was high. The data that led to an unconformity were re-examined and updated.

### *Data Analysis*

The study data were analyzed via descriptive analysis. Descriptive analysis refers to summarizing the data in accordance with the previously determined themes, and it consists of four stages. At the first stage, a framework is created for the analysis. In this line, the first stage of this study was to review literature and create a first form based on expert opinions. The second stage requires filling the data in the thematic framework. In the current study, the data in the 39 articles mentioned above were coded in line with the items included in the forms created at the first stage. The third and fourth stages of a descriptive analysis are describing and interpreting the data respectively. Within the framework of the current study, the data that were processed in line with the last two stages of a descriptive analysis were described and interpreted receiving support by the details in the coded articles (Yıldırım & Şimşek, 2013). Moreover, the data gathered from the articles were coded on Microsoft Excel and frequency/percentage were calculated.

## Results

### Type of Items Used in the Test

When the published articles were reviewed to see the type of items, it was clear that achievement tests used in 13 (33,33%) of the reviewed articles included multiple-choice tests, 21 (53,85%) of the articles included tests with open-ended items, one (2,56%) of them included a test with items of mixed type. Furthermore, achievement was measured through a diagnostic branched tree (DBT) in one of the articles (2,56%). On the other hand, the items in mixed-type tests were open-ended and multiple-choice. Moreover, three (7,69%) of the coded articles did not specify the type of the items in the test.

### The Test Theory Used in the Process

The researchers also examined on which test theory were based for developing the achievement test. The analysis showed that 36 (92,31%) of the 39 articles included achievement tests relying on classical test theory (CTT). Among these 36 articles, one of the articles additionally took the advantage of item response theory (IRT), while one other additionally relied on Rasch analysis, which is based on IRT. Two (5,13%) of the 39 articles included tests that were developed via analysis depending on IRT. One (2,56%) of the articles relied on individualized computer-based test.

### Using Grading Key for the Tests with Open-Ended Items

When the articles were examined to see the use of grading keys, it was clear that out of 21 articles that included open-ended test items, 13 (61,90%) of them developed a grading key, while one (4,76%) of them used a ready-made grading key that was developed within the scope of PISA. Seven (33,33%) of the articles either did not use a grading key or specify if a grading key was developed. The researchers also examined the type of the grading keys in 13 articles that developed a rubric. In this line, six (46,15%) of the articles used a holistic rubric. In the light of this, it is possible to say that a holistic rubric was used in grading in general. Moreover, three (23,08%) of the articles stated to have used partial grading. Additionally, two (15,38%) of the articles conducted grading in line with the dimensions of the taxonomy, which was used to develop the achievement test, while an answer key was prepared in one (7,69%) article, and one (7,69%) of the articles used a rating scale.

The researchers also examined if the 13 articles which developed a rubric also included the rubric itself or an example of it. In this line, it was seen that eight (61,54%) of the articles included the rubric, whereas five (38,46%) of them did not include the developed rubric.

### Number of Items in the Draft and Final Form

Out of 38 articles in which the type of items was not DBT, four (10,53%) of them did not specify the number of items in the draft form. The article which used a computerized adaptive testing (CAT) that the number of items in the draft item pool was 838. It was clear that the number of items in the draft form varied between 3 and 124 in other 33 articles. When the researchers examined the number of items in the final forms, one (2,56%) of the articles did not specify the number of items in the final form. In the article which relied on CAT, there were 752 items in the final item pool, the number of items in the form applied to the students was 80. In the article which measured achievement in maths with DBT, the test included five items which the students were supposed to respond. The number of items in the remaining 36 articles varied between 3 and 80. Also, the number of items in the draft form and final form was equal in 21 (53,85%) of the articles.

### *Including Test Form, Examples of Items and Instruction*

When the articles were examined to see if the articles included the test forms, it was clear that there were eight (20,51%) articles that included the whole test form, while one (2,56%) article also included all the items in the test as it organized the findings in terms of each test item. However, 30 (76,92%) of the articles did not include the test form. On the other hand, none of the eight articles that included the test form had an instruction (0,00%). Additionally, there was one (3,33%) article that included an instruction although it did not include the test form. When the articles were examined to see if they included an example item, 12 (40,00%) of the 30 articles that did not include the whole test form included examples of items. Out of these 12 articles, one (8,33%) of them included all the items (3 items). 18 (60,00%) articles did not even include an example item.

### *Including the Test Goal*

When the articles were examined to see if they specified the test goal or not, it was seen that 36 (92,31%) of the 39 articles reported the test goal. On the other hand, three (7,69%) of them did not include the test goal. When the researchers examined the articles to see if they reported for which subject the test was developed, they noticed that 34 (87,18%) of the articles reported about the subject, five (12,82%) of them did not report it. On the other hand, the analysis showed that the study group at which the tests were used were reported in all the articles. Table 2 shows for which study group the tests were developed.

Table 2. Frequency and percentage distribution of study groups for the developed maths achievement tests

| **Study group** | **The number of articles (f)** | **The percentage of articles (%)** |
|---|---|---|
| Teachers | 1 | 2,56 |
| University | 7 | 17,95 |
| 9th, 11th and 12th Grade | 6 | 15,39 |
| 5th, 6th, 7th and 8th Grade | 22 | 56,41 |
| 1st, 3rd and 4th Grade | 3 | 7,69 |
| Total | 39 | 100 |

As is seen in Table 2, the highest number of maths achievement tests were developed at secondary school level. There was only one article that developed an achievement test for teachers.

### *Results Regarding the Table of Specifications*

Table 3 shows the frequency and percentage distribution for including table of specifications in the article, receiving expert opinions in articles with a table of specifications, and the information on the taxonomy used in the table of specifications.

Table 3. Frequency and percentage distribution for table of specifications and receiving expert opinions

| **The status of prepared table of specification** | |
|---|---|
| Articles that prepared table of specification or give information about subject/objectives and item distribution | Articles that did not prepare or give any information on the table of specifications |
| 9(%23,08) | 30(%76,92) |
| **The status of received expert opinion on the prepared table of specifications.** | |
| Received | Did Not Receive or Give Any Information |
| 2(%22,22) | 7(%77,78) |
| **The status of taxonomies that used to prepare the table of specifications** | |
| Informed | Did Not Inform |
| 3(%33,33) | 6(%67,77) |

As is seen in Table 3, most of the articles included in the current study did not give any information on the table of specifications (f=30; 76,92%). On the other hand, there were nine (23,08%) articles that prepared a table of specifications or included details such as subject/objectives and item distribution. Also, there were only two (22,22%) articles that received expert opinion on the prepared table of specifications. When the researchers examined the qualifications and the number of experts whose opinions were received about the table of specifications, it was clear that one of the articles obtained opinions of three classroom teachers, three academicians in the field of maths teaching, one academician in the field of educational measurement and assessment, one academician in the field of curriculum and instruction, whereas the other article received opinions from one academician in the field of maths teaching and four maths teachers. When it comes to the taxonomy used to prepare the table of specifications, six (66,67%) of the articles did not include an information on taxonomy, three (33,33%) of the articles specified the taxonomy that was used. Out of the three articles that specified the used taxonomy, one of them stated to have used the criteria of Dienes principles, while one other relied on Bloom taxonomy. One other article specified that the items were prepared in line with the levels of graph interpretation by Curcio (1987).

### *Results Regarding the Item Pool*

Table 4 shows the results on creating an item pool, receiving expert opinions for the items in the pool, specifying the resources which were used to write items and the use of previously created ready items.

Table 4. Frequency and percentage distribution for item pool, expert opinion and use of previously created ready items

| | **Yes** | **No/Did Not Informed** |
|---|---|---|
| Creating | 10(%25,64) | 29(%74,36) |
| Receiving expert opinion | 7(%70,00) | 3(%30,00) |
| Giving information about used sources | 18(%46,15) | 21(%53,85) |
| Using ready items | 13(%33,33) | 26(%67,67) |

As is seen in Table 4, out of 39 articles, an item pool was created in 10 (25,64%) articles. Seven of these articles received expert opinions. Furthermore, in one of the seven articles, it was stated that the items were created by experts themselves. The number of experts varied between 2 and 15 in these seven articles. When the researchers examined in which fields the experts were studying, they found out that the experts were studying in the fields of maths teaching, maths teachers, classroom teachers, educational assessment, and evaluation as well as curriculum and instruction. Another feature about the experts was that there was no expert of language.

As is seen in Table 4, when it comes to specifying the resources that were used to write items, the resources were specified in 18 articles, while 21 of the articles did not include any information about the resources. Furthermore, considering creating item pool, the researchers found out that 26 of the articles did not use any previously created ready items. The researchers also examined how many ready items were used in the articles that made use of ready items. In two of these 13 articles, there was no information about the number of ready items used in the article. In three of the articles, some of the items were previously written, whereas all the items were ready in the remaining eight articles. The number of ready items used in the articles varied between 2 and 124.

### *Results Regarding Pilot Testing*

When the researchers examined whether a pilot testing was conducted, they found out that there were 17 (43,59%) articles that did not give any information if a pilot testing was conducted or not, while there were 18 (46,15%) articles that conducted a pilot testing. In two (5,13%) of the remaining four articles, it was stated that a selection was made by using the statistics of the ready items. Moreover, one (2,56%) of the articles itself was a pilot testing, while one (2,56%) of them conducted a pilot testing in another study earlier. When the researchers examined the number of participants with whom a pilot testing was conducted, they concluded that five (27,78%) of the articles did not give any information about that. In 13 (72,22%) articles that gave information about that, the number of participants varied between 15 and 1130. Table 5 shows the frequency and percentage distribution for the analysis conducted in articles that used a pilot testing.

Table 5. Frequency and percentage distribution for the analysis conducted during pilot testing

| | **Yes** | **No/Did Not Informed** |
|---|---|---|
| Distractor Analysis (MCI=7) | 1(%14,29) | 6(%85,71) |
| Item Discrimination | 8(%44,44) | 10(%55,56) |
| Item Difficulty | 6(%33,33) | 12(%66,67) |
| Reliability | 11(%61,11) | 7(%38,89) |
| Factor Analysis | 0(%0,00) | 18(%100,00) |
| Others | 0(%0,00) | 18(%100,00) |

As is seen in Table 5, out of 14 articles that used multiple-choice item, seven articles conducted a pilot testing, and a distractor analysis was done only in one (14,29%) of the articles. However, no distractor analysis was conducted in the other six articles. The article that conducted a distractor analysis did not specify the program with which the analysis was done. Considering the findings about item discrimination in pilot application, the number of articles that examined item discrimination (f=8; 44,44%) was lower than the number of articles that did not (f=10; 55,56%). On the other hand, the two (25,00%) articles that examined item discrimination did not give any information about the statistics with which the analysis was conducted. Six (75,00%) articles that gave information about item discrimination made use of one or a few of item discrimination index, item-total correlation, item-remainder correlation, upper group-lower group discrimination and independent groups T test results. Moreover, none of the articles specified the program with which discrimination was calculated. The researchers also found out that the number of articles that examined item difficulty in the pilot testing (f=6; 33,33%) was lower than the number of articles that did not (f=12; 66,67%). Moreover, the articles that examined item difficulty after the pilot testing also examined item discrimination. In this case, in two (11,11%) articles, item difficulty was not examined, while item discrimination was examined. Additionally, none of the articles specified the program with which item difficulty was calculated.

As is seen in Table 5, which shows the findings on examining reliability coefficient in the pilot testing, there were 11 (61,11%) articles that examined reliability, while there were seven (38,89%) articles that did not. When the type of reliability coefficient was examined, out of 11 articles, three of them (27,27%) calculated Cronbach alfa, six of them (54,55%) calculated KR-20, one of them (9,09%) calculated KR-21 and one of them (9,09%) calculated rater reliability. There was only one (9,09%) article that gave information about the program with which reliability was calculated, and the program was stated to be SPSS. When the researchers examined the findings about the factor analysis in the pilot testing, it was seen that none of the articles (f=18; 100,00%) conducted factor analysis. Lastly, when the articles were examined to see if they included any other analysis within the scope of the psychometric qualities of the test, it was clear that none of the articles (f=18; 100,00%) included additional analysis.

Table 6 shows the frequency and percentage distribution for the analysis at the end of the pilot testing.

Table 6. Frequency and percentage distribution for the changes in the item pool at the end of the pilot testing

| | **Yes** | **No/Did Not Informed** |
|---|---|---|
| As a result of the pilot test, were any changes/improvements made in the items? | 3(%16,67) | 15(%83,33) |
| Were items removed from the test as a result of the pilot test? | 8(%44,44) | 10(%55,56) |
| If items were removed, did any information give about these items how affect content validity? | 0(%0,00) | 8(%100,00) |
| After the pilot test, was the expert opinion taken to remove the item or to make changes/improvements? | 2(%11,11) | 16(%90,09) |

As is seen in Table 6, there are three (16,67%) articles which made changes/improvements at the end of the pilot testing. When these three articles were examined, it was clear that none of them reported how many items were changed. However, in one of them, it was stated that the changes made to the items were about language and mechanics. Also, there are two (5,13%) other articles that made changes/improvements depending on the expert opinions although they did not conduct a pilot testing. These articles did not give any information about the number of items that were changed. When the researchers examined the experts that contributed to changes/improvements in the items, it was seen that in one of articles, opinions of three maths teachers and three academicians were received, while in the other article, two academicians gave opinion. Considering the removal of items at the end of the pilot testing, there were eight (44,44%) articles that removed items, whereas there were 10 (55,56%) articles that did not remove any items from the tests. None of the eight articles in which an item was removed reported how removing the items affected content validity. Moreover, in these eight articles, the number of items removed from the tests varied between 1 and 38.

When the researchers examined the articles to see if expert opinion was received to make any changes/improvements in the items or remove items after the pilot testing, they found out that only two (11,11%) of the 18 articles received expert opinion. In the remaining 16 articles, no expert opinion was received even though there was a change/improvement in the items and item pool, which means that the changes/improvements were not based on expert opinions. In one of the two articles that received expert opinions, there was no information about the experts, whereas the experts whose opinions were received were an academician and two BILSEM teachers in the other article.

### *Results About the Real Study*

Table 7 shows the frequency and percentage distribution for the analysis conducted in the real study.

Table 7. Frequency and percentage distribution for the analysis conducted in the real study

| | **Yes** | **No/Did Not Informed** |
|---|---|---|
| Distractor Analysis (MCI=14) | 1(%7,14) | 13(%92,86) |
| Item Discrimination | 10(%25,64) | 29(%74,36) |
| Item Difficulty | 8(%20,51) | 31(%79,49) |
| Reliability | 22(%56,41) | 17(%43,59) |
| EFA/CFA | 4(%10,26) | 35(%89,74) |
| Others | 6(%15,38) | 33(%84,62) |

As is clear in Table 7, out of 14 articles that used a multiple-choice test including the mixed tests, distractor analysis was conducted only in one (7,14%) article. This article did not specify the program with which distractor analysis was conducted. Considering the findings about item discrimination in the real study, the number of articles that examined item discrimination (f=10; 25,64%) was lower than the number of articles that did not (f=29; 74,36%). In five (50,00%) articles that examined item discrimination, there was no information about which statistics was used to calculate item discrimination. In five (50,00%) articles that specified the statistics used to examine item discrimination, the specified statistics included item-total correlation, upper group-lower group discrimination, *a* parameters and multifaceted Rasch.  In terms of item discrimination, the programs of MULTILOG 7.03 and IRTPRO were used in two (33,33%) articles that predicted a parameter depending on IRT, the program called FACETS was used to make a prediction in one (16,67%) article that relied on multifaceted Rasch. Moreover, in one (16,67%) article, the program TAP was used. In the other two (33,33%) articles, the program that was used to calculate item discrimination was not specified.

As is seen in Table 7, the number of articles that examined item difficulty in the real study (f=8; 20,51%) was lower than the number of articles that did not (f=31; 79,49%). Moreover, item discrimination was examined in all articles that examined item difficulty after the real study. Also, item difficulty was not examined although item discrimination was examined in two (5,13%) articles. Considering the programs with which item difficulty was calculated, the programs of MULTILOG 7.03 and IRTPRO were used in two (66,67%) articles that predicted *b* parameter depending on IRT, whereas the program called TAP was used in one (33,33%) article.

Considering the findings on reliability in the real study, the number of articles that calculated reliability coefficient (f=22; 56,41%) was higher than the number of articles that did not (f=17; 43,59%). When the researchers examined the reliability coefficients reported in the articles, they found out that the articles used one or more of coefficients including Cronbach alfa, Cohen kappa, KR-20, reverse of standard deviation, reliability based on IRT, multifaceted Rasch reliability index, Spearman-Brown split half, Kendall coefficient of concordance. Moreover, in some of the articles, the expression of coder/rater reliability was used although the type was not specified. When the researchers reviewed the articles that reported reliability coefficient to see if they stated why the related reliability coefficient was chosen, they concluded that only three (13,64%) of the articles specified the reason of the choice. Lastly, considering the findings about the program with which reliability coefficient was calculated, 18 (81,82%) articles did not specify the program used. In the remaining four

(4,55%) articles, the programs used were respectively FACETS, MULTILOG 7.03, SPSS and TAP.

When the researchers examined the articles to see if they conducted EFA/CFA or not, they found out that a factor analysis was conducted in four (10,26%) articles. EFA was conducted in three (75,00%) of these articles, while a factor analysis was done but the type was not specified in one (25,00%) of them. At the end of EFA, two of the articles were found to be one dimensional, while one of the articles was found to be six dimensional. In the article that did not specify the type of the factor analysis, there was five one-dimensional sub-tests. Moreover, the three articles that conducted EFA gave information about the amount of variance, and the variance explained varied between 30,84% and 53,59%. On the other hand, there was only one (33,33%) article that reported the program with which EFA was conducted. It was seen that none of the articles conducted CFA. Considering the use of other analysis within the scope of the psychometric qualities of the test, 33 (84,62%) of the articles did not include any additional analysis, while additional analysis was conducted in six (15,38%) of the articles. The statistics of these analysis included mean and standard deviation of the items, correlation between the item difficulty of parallel forms and the significance of the average difference between the forms, test discrimination, average point biserial correlation value as well as skewness and kurtosis coefficients. Moreover, in one of these six articles, analysis was conducted based on CAT. Considering the information on the program that was used in the articles including additional analysis, there was only one (16,67%) article that stated that the analysis was conducted via SPSS, while the other five (83,33%) articles did not give information about the program. Considering the number of participants in the real study, this number varied between 3 and 865.

### *Results about Reliability and Validity*

The researchers examined the articles to see if they reported reliability coefficient separately in pilot testing and real study. No matter if it was in the pilot testing or real study, there were 33 (84,62%) articles that reported reliability coefficient. On the other hand, six (15,38%) of the articles did not report reliability coefficient in either pilot testing or real study. In this line, it is possible to say that the number of articles that reported reliability coefficient was higher than the number of articles that did not. Out of 33 articles that reported reliability coefficient, there were 11 (33,33%) articles that reported reliability coefficient only in the pilot testing, while there were 20 (60,61%) articles that reported reliability coefficient only in real study. Moreover, two (6,06%) of the articles reported reliability coefficient both in the pilot testing and real study.

Table 8 shows the frequency and percentage distribution for the types of validity.

Table 8. Frequency and percentage distribution for the types of validity

| | **Yes** | **No/Did Not Informed** |
|---|---|---|
| Content Validity | 23(%58,97) | 16(%41,03) |
| Construct Validity | 6(%15,38) | 33(%84,62) |
| Criterion-Related Validity | 3(%7,69) | 36(%92,31) |

Out of 39 articles that developed a maths achievement test, there was evidence of content validity in 23 (58,97%) articles. On the other hand, there was no evidence of content validity in 16 (41,03%) articles. Considering the evidence regarding content validity in these 23 articles, 14 (60,87%) articles received expert opinions, eight (34,78%) articles both received expert opinions and prepared a table of specifications, one (4,35%) article only prepared a table of specifications. Out of 22 articles that received expert opinions, six (27,27%) of them did not report the number of experts, while in four (18,18%) of these six articles, the qualifications of the experts were not specified. On the other, the qualifications and number of experts were reported in 16 (72,73%) articles. When the researchers examined the qualifications of the experts whose opinions were received in the articles that reported the number and qualifications of experts, the experts were classroom teacher, maths teacher, academician, expert of language, expert of educational measurement and evaluation, expert of curriculum and instruction, expert of educational sciences or BİLSEM teacher. When it comes to the stage at which expert opinions were received, expert opinions were received while examining the item pool, after the pilot testing, after preparing the items, while choosing the items and before the final selection. Two (9,09%) of the articles did not specify at which stage expert opinions were received.

When the findings about the construct validity were examined, it came out that there was no evidence of construct validity in 33 (84,62%) of the articles, while six (15,38%) of them provided evidence of construct validity. Examining the evidence of construct validity showed that factor analysis was done in three (50,00%) articles, model-data fit was reported in one (16,67%) article, a factor analysis was done, and model-data fit was reported at the same time in one (16,67%) article, inter-item correlation was examined in one (16,67%) article. Moreover, in one of the articles that conducted factor analysis, it was stated that internal validity would be high as PISA items were used in the study. In one other article that examined inter-item correlation, having items that could not be solved through memorization due to being real-life problems was provided as evidence of validity.

When the researchers examined the findings regarding the criterion-related validity, there was not any evidence of criterion-related validity in 36 (92,31%) articles, while three (7,69%) articles provided evidence of criterion-related validity. Two articles used end-of-the-year and end-of-the-term grades to calculate correlation with the scores obtained from the achievement test, which means that they made use of convergent criterion validity. In the other article, the researchers did the same analysis for the achievement test and course grades and examined if the same results were obtained or not.

### *Results About the Setting of Test Administration*

When the findings about the setting of test administration were examined, it was clear that the setting or conditions were not specified in 25 (64,10%) articles. There was information about the conditions of test administration only in 14 (35,90%) articles. The conditions about which information was provided included administering the test on computer, duration of the test, the setting, how many days it took to administer the test, use of optic forms, providing test operators with an instruction and administering the test in line with these conditions.

## Discussion, Conclusion and Suggestions

According to the study findings, the tests developed in the articles mostly included open-ended items that made it possible for students to construct their own answers. It was clear that more than half of the articles that made use of open-ended items developed a rubric. Most of the articles that developed a rubric used a holistic rubric. The reason why a holistic rubric was preferred in most articles might be that a holistic rubric enables a holistic evaluation, which means that it does not require separate criteria and so is more practical (Brookhart, 2018; Hunter et al., 1996). However, using an analytical rubric instead of a holistic one to score open-ended questions can be recommended to test developers as analytical rubrics enhance reliability in terms of test sensitivity, they address student success in line with previously-determined criteria in a more detailed way, they are more likely to reveal which criteria the students master or fail, and they can give more accurate feedback (Brookhart, 2018; Goodrich-Andrade, 2000, 2001; Moskal, 2000; Mertler, 2000; Şahin, 2019). Similarly, there are primary studies as well as reviews that conclude that analytical rubrics measure with a higher level of reliability when compared to holistic rubrics (Büyükkıdık, 2012; Jonsson & Svingby, 2007; Öksüzoğlu, 2022; Reznitskaya et al., 2009; Şanlı, 2010; Yıldıztekin, 2014). In addition to that, the researchers found out in the current study that one-third of articles that used open-ended items did not develop a rubric or did not give any information about that even if they did. Open-ended items are subjective in terms of scoring, which makes it more important to prepare a rubric in order to increase validity and reliability, because preparing a rubric will increase the objectivity in scoring (Goodrich-Andrade, 2005; Moskal, 2000; Moskal & Leydens, 2000). Moreover, it can be recommended to report about the process of preparing a rubric and present an example rubric at least in order to reach transparency for test administers, researchers or readers.

It was concluded that the tests developed were generally developed according to CTT. CTT is a theory that is frequently preferred by educators in the field in order to develop a test, obtain item and test statistics. The reason why this theory is mostly preferred might be that measurement and administration is easy when the test is based on this theory, most researchers including graduates are competent in this theory, it is easy to reach and use analysis programs developed in line with this theory. The other reason might be that test developers do not know about IRT. Also, the researchers might have resorted to CTT because the number of individuals that must be reached for item parameters prediction is high according to IRT, the process of developing a test according to IRT is complex, educators are generally not familiar with this theory and analysis programs and they might have assumptions (Crocker & Algina, 2006; De Ayala, 2009). However, when some conditions are met such as ensuring the assumptions, choosing a model that fits the data and reaching the appropriate sample size, IRT has the advantage of parameter invariance (predicting item and individual parameters separately) (DeMars, 2010; Doğan & Kılıç, 2017; Hambelton & Swaminathan, 1985; Hambleton et al., 1991, Lord & Novick, 1968). Moreover, IRT has another advantage of predicting item and individual parameters more accurately if the aforementioned conditions are ensured. When the studies that compared test development in accordance with CTT and IRT are reviewed, there are some studies in the literature that conclude that when the necessary conditions are met, there is no difference between the results obtained from the two theories, while there is a high relation between the scores and discrimination obtained from them (Akyıldız & Şahin, 2017; Çelen, 2008; Çelen & Aybek, 2013, Mor-Dirlik, 2021). However, there are some other studies that underline that the scores obtained from the two theories are not interchangeable (Akyıldız & Şahin, 2017). In the light of this, test developers can be recommended to examine the assumptions/conditions regarding the theories first of all, and then decide on the test theory on which they will develop a test.

It was found out at the end of the current study that almost all the articles examined within the scope of the current study reported the goal of the test, while all of them gave information about the study group which the test targets. Unlike the current study findings, Karadağ (2011) carried out a study to examine dissertations and found out that almost none of the studies examined within the scope of that study reported the reason why the related tests were developed. It is important to report the goal of the tests as this will give an idea to test administers and researchers/readers about what the test aims to measure, the fields where the test can be used, difficulty of the test or if the content of the test complies with its goal (AERA et al., 2014), and this will be an evidence of validity regarding the measurements obtained from the test (Lane et al., 2016). Moreover, considering validation of a construct, the most important step is to define the construct (Tindal & Haladyna, 2012). As to the inclusion of test form or example items in the study, it was concluded that most of articles did not include the test form. Most of the articles that did not include the test form did not include any example items, either. This study finding is supported by the study conducted by Karadağ (2011), who found out that almost none of the dissertations included example items. Researchers can be recommended to include example items, if not the whole test form, in studies that develop an achievement test. It is important to include the whole test form or example items in the related study for the sake of effective feedback to the related individuals and institutions in terms of the conformity between the test goal and test content (AERA et al., 2014). Additionally, including the whole test form or example items can give an idea to test administers and researchers/readers about the content, level and understandability of the achievement test.

Lane et al. (2016) state that designating the test content depends on writing effective test items and expert opinions. According to this, creating an item pool is an important step of test development process, while it was concluded at the end of the current study that most of the articles did not create an item pool. It was seen that most of the articles that created an item pool received expert opinions regarding the item pool. When the qualifications of these experts were examined, it was clear that the experts had different areas of expertise. Besides experts within the field, the articles received the opinions of experts studying in the field of educational measurement and evaluation, educational curriculum and instruction as well as teaching professionals, which can be considered as an evidence of validity. Furthermore, it was concluded that none of the articles received opinions of a language expert at this step. It is important to receive opinions of a Turkish language expert at this stage in order to ensure the appropriateness of the test to the level of students as well as its legibility and understandability. Because of that reason, test developers are recommended to receive expert opinions from Turkish language experts.

The quality of items is generally assessed via examining the items and trying them, which is called as pre-testing. At this stage, the items are reviewed in terms of the content quality, clarity and variables that affect the responses of test takers but are not related to the construct that the test is trying to measure (AERA, et al., 2014). In this line, conducting a pre-testing after creating the item pool is important to calculate item and test statistics so as to choose the items and make the necessary changes/improvements, specify the setting and conditions to administer the test as well as the points that should be changed with the test form. In short, it is possible to identify the applicability of the test and find out how it can be improved by means of pre-testing (Enago, 2021). It was found out at the end of the current study that half of the articles that were examined within the scope of the current study did not conduct a pre-testing or did not specify even if they did. Furthermore, a pilot testing can be conducted with a small sample group similar to the sample of the real study or interviews can be conducted about the items in order to identify if the items in the test are understandable (Haladyna & Rodriguez, 2013). In this line, test developers can be recommended to conduct pre-testing. Additionally, Karadağ (2011) conducted a study which examined the pilot-testing process and its results in developing an achievement test and found out that almost none of the studies examined report about the process and results of pre-testing. It was also concluded that most of the articles that conducted a pre-testing did not examine the statistics of item discrimination and item difficulty as an evidence of construct validity or they did not report about the results even if they examined the related statistics. Moreover, considering the articles that conducted pre-testing, the number of articles that reported about test reliability was higher than the number of those that did not. Also, the articles that reported reliability coefficient mostly calculated KR-20 coefficient. Similarly, Mutluer and Yandı (2012) carried out a study that examined dissertations and concluded that the most frequently reported reliability coefficient is Cronbach alfa and KR20/21 coefficients. When the articles were reviewed to see if factor analysis was conducted in pre-testing, it was found out that none of the articles that conducted a pre-testing performed factor analysis. It was also revealed that any improvement/change was not performed in the item pool after the pre-testing in most of the articles, or they did not receive expert opinions to make a change/improvement even if they did. Besides, at the end of the pre-testing, some items were removed from the item pool in almost half of the articles although none of these articles specified how the removed item affected the content validity of the items. While deciding on removing or changing/improving an item based on the analysis during the pre-testing, it is of vital importance to carry out this process not only depending on statistics but also after receiving expert opinions on how content validity will be affected.

The study findings show that only one of the articles that used multiple-choice test items in the real study made distractor analysis. It was also found out that item discrimination and item difficulty statistics, which are important evidence of validity, were not calculated or reported for the real study in most of the articles. It was seen that more than half of the articles reported test reliability coefficient during the real study. The articles that reported reliability coefficient mostly calculated Cronbach alfa coefficient. Unlike the current study findings, Evrekli et al. (2011) found out that in the dissertations examined within the framework of their study, the reliability process of the data collection tools was generally insufficient or partly sufficient. Also, Karadağ (2011) concluded that there were some mistakes about identifying the reliability.

When the results about reliability are considered in general, it is clear that the number of articles that reported reliability coefficient in the real study was higher than those that reported it in pre-testing. However, reliability coefficient was reported neither in pre-testing nor in real study only in two articles. On the other hand, six articles reported reliability coefficient both in pre-testing and real study. When the results about validity are considered in general, it is obvious that most of the articles provided evidence of content validity, whereas almost none of the articles provided evidence of criterion validity. Moreover, when the evidences regarding the construct validity were examined in terms of factor analysis, model-data fit and inter-item correlation, most of the articles did not provide any of these evidences. This finding is supported by previous study findings in the literature. Evrekli et al. (2011) also concluded that the validity processes of the dissertations examined in their study were insufficient in general or partly sufficient. Similarly, Karadağ (2011) found out that there were some mistakes about identifying validity. In the light of the current study finding and previous study findings in the literature, test developers can be recommended to diversify evidences of validity regarding the tests they develop, examine if they measure in line with the test goal or not, and report the related results. Considering combined construct validity based on collecting evidence (Messick, 1995), it is very important to collect evidence in test development process, while providing evidence not only for content validity as a traditional validity, as is the case in the articles examined in the current study, but also for construct and criterion validity will strengthen the article methodologically. Moreover, as suggested by Cronbach (1989), unless construct validity is well-determined, it is prone to be too open-ended and it can lead to an excessively long or even infinite process (cited by Lane, 2016).

It was concluded in the current study that the evidence of content validity that was most-frequently used was receiving expert opinions. However, it was found out that more than one-third of the articles did not receive expert opinions. Moreover, the qualifications and number of experts whose opinions were received were not specified in the articles that resorted to expert opinions. The current study finding was supported by Evrekli et al. (2011), who found out that the dissertations they examined did not receive expert opinions while developing a scale or test. Haladyna and Rodriguez (2013, p. 322) also indicate that it is necessary to resort to experts who will evaluate items and types of responses in developing items, and to report the demographic qualities, qualifications, and experiences of these experts in the study. In this line, test developers and researchers can be recommended to receive opinions of experts from various fields (the related field, language, assessment and evaluation, curriculum and instruction, etc.) in order to obtain valid measurements. In a few of the articles that were examined in the current study, a table of specifications was prepared as an evidence of content validity. Similarly, Evrekli et al. (2011) concluded in their study that a table of specifications was not prepared or was prepared inaccurately. Moreover, Boyraz (2018) conducted a study to examine dissertations and concluded that the analysis for the sake of content validity and the level of preparing a table of specifications were not sufficient.

Lastly, when the setting of test administration was examined, it was seen that most of the articles did not give any information about how to administer the test, test duration or setting. Likewise, Karadağ (2011) stated that almost none of the dissertations examined in the study made explanations about how to score achievement tests and assessment tools, and especially about interpreting the scores obtained from data collection tools. It is important to report the setting and conditions to administer the test for those who will administer the test in the future and researchers, and so researchers are recommended to report on that.

The current study aims at examining the achievement tests developed in the articles published between the years of 2015 and 2020 in the journals indexed in SSCI or ESCI in line with the previously determined criteria. Similar studies can be conducted in the future with an updated set of articles. Additionally, researchers can include journals of different indices or other articles indexed in the same journals in future studies. Apart from that, they can update and deepen the criteria. Lastly, future studies can be conducted in a different field from Maths, or with dissertations instead of articles as is the case in the current study.

### *Ethics Committee Approval*

Hacettepe University Ethics Committee was applied for the approval of the ethics committee for the study. Ethics committee found the study ethically appropriate and ethics committee approval document (dated 05.05.2022 and numbered E-35853172-755.02.05-00002166209) was received.

Note: A part of this study was presented as an oral presentation at Izmir Democracy University 1st International Educational Research Congress (IDU-ICER'19).

## References

Acar-Güvendir, M., & Özer-Özkan, Y. (2015). The examination of scale development and scale adaptation articles published in Turkish academic journals on education. *Electronic Journal of Social Sciences, 14*(52), 23-33. doi: 10.17755/esosder.54872

AERA, APA, & NCME. (2014). *Standarts for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Boztunç-Öztürk, N. B., Eroğlu, M. G., & Kelecioğlu, H. (2015). A review of articles concerning scale adaptation in the field of education. *Education and Science, 40*(178), 123-137. doi: 10.15390/EB.2015.4091

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22), 1-12. doi: 10.3389/feduc.2018.00022.

Büyükkıdık, S. (2012). *Comparison of interrater reliability based on the classical test theory and generalizability theory in problem solving skills assessment*. (Published master thesis). Hacettepe University, Ankara.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Ohio, Maison: Cengage Learning.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5. ed.). New York, NY: Harper & Row Publishers Inc.

Çelen, Ü. (2008). Comparison of validity and reliability of two tests developed by classical test theory and item response theory. *Elementary Education Online, 7*(3), 758-768. Retrieved from https://dergipark.org.tr/en/download/article-file/90935

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology, 4*(2), 64-75. Retrieved from https://dergipark.org.tr/en/download/article-file/65958

Çetin, B. (2019). Test geliştirme. B. Çetin (Ed.). In *Eğitimde ölçme ve değerlendirme [Measurement and assessment in education]* (p. 105-126). Ankara: Anı Publishing.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

Delice, A., & Ergene, Ö. (2015). Investigation of scale development and adaptation studies: An example of mathematics education articles. *Karaelmas Journal of Educational Sciences, 3*(1), 60-75. Retrieved from https://dergipark.org.tr/tr/pub/kebd/issue/67216/1049114

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Doğan, N., & Kılıç, A. F. (2017). Madde tepki kuramı yetenek ve madde parametre kestirimlerinin değişmezliğinin incelenmesi. Ö. Demirel and S. Dinçer (Eds.). In *Küreselleşen dünyada eğitim [Education in a globalizing world]* (p. 298-314). Ankara: Pegem Academy. doi: 10.14527/9786053188407.21

Downing, S. M., & Haladyna, T. M. (2011). *Handbook of test development.* New Jersey, NJ: Lawrence Erlbaum Associates Publishers.

Enago (2021). *Why is a pilot study important in research?*. Retrieved from https://www.enago.com/academy/pilot-study-defines-a-good-research-design/

Ergene, Ö. (2020). Scale development and adaptation articles in the field of mathematics education: Descriptive content analysis. *Journal of Education for Life, 34*(2), 360-383. doi:10.33308/26674874.2020342207

Evrekli, E., İnel, D. , Deniş, H., & Balım, A. G. (2011). Methodological and statistical problems in graduate theses in the field of science education. *Elementary Education Online, 10*(1), 206-218. Retrieved from https://dergipark.org.tr/tr/pub/ilkonline/issue/8593/106858

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3. ed.). New Jersey, NJ: Lawrence Erlbaum Associates Publishers.

Goodrich-Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13-18. Retrieved from https://eric.ed.gov/?id=EJ609600

Goodrich-Andrade, H. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4), 1-22. Retrieved from https://cie.asu.edu/ojs/index.php/cieatasu/article/view/1630

Goodrich-Andrade, H. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27-31. doi: 10.3200/CTCH.53.1.27-31

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and Applications*. Dordrecht, The Netherlands: Kluwer-Nijhoff Publishing Co.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). California, CA: Sage.

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, *11*(2), 61-85. Retrieved from https://www.evaluationcanada.ca/secure/11-2-061.pdf

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

Karadağ, E. (2011). Instruments used in doctoral dissertations in educational sciences in Turkey: Quality of research and analytical errors. *Educational Sciences: Theory & Practice*, *11*(1), 311-334. Retrieved from https://silo.tips/download/eitim-bilimleri-doktora-tezlerinde-kullanlan-lme-aralar-nitelik-dzeyleri-ve-anal

Lane, S., Raymond, M. R., & Haladyna, T. M. (2016). *Handbook of test development* (2. ed.). New York, NY: Routledge.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Mertler, C.A. (2000). Designing scoring rubrics for your classroom. *Practical Assessment, Research, and Evaluation*, *7*(25), 1-8. doi: 10.7275/gcy8-0w24

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749. doi:10.1037/0003-066x.50.9.741

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2. ed.). Thousand Oaks, CA: Sage.

Mor-Dirlik, E. (2014). The analysis of the doctoral dissertations themed of scale development according to the test and scale development standards. *Journal of Measurement and Evaluation in Education and Psychology, 5*(2), 62-78. doi: 10.21031/epod.63138

Mor-Dirlik, E. (2021). The comparison of item discrimination parameters estimated from different test theories. *Trakya Journal of Education. 11*(2), 732-744. doi: 10.24315/tred.700445

Moskal, B. M. (2000). Scoring rubrics: What, when and how?. *Practical Assessment, Research, and Evaluation*, *7*(3), 1-5. Doi: 10.7275/a5vq-7q66

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research, and Evaluation*, *7*(4), 1-22. doi: 10.7275/q7rm-gg74

Mutluer, C., & Yandı, A. (2012, September). Türkiye'deki üniversitelerde 2010-2012 yılları arasında yayımlanan tezlerdeki başarı testlerin incelenmesi [The examination of achievement tests in theses published in universities in Turkey between 2010-2012]. Paper presented at the *Eğitimde ve Psikolojide Ölçme ve Değerlendirme III. Ulusal Kongresi*, Turkey: Bolu. Abstract retrieved from https://www.epodder.org/wp-content/uploads/2020/07/EPOD-2012.pdf

Öksüzoğlu, M. (2022). *The investigation of items measuring high-level thinking skills in terms of student score and score reliability.* (Unpublished master thesis). Hacettepe University, Ankara.

Özçelik, D. A. (1992). *Ölçme ve değerlendirme [Measurement and assessment].* Ankara: ÖSYM Publ.

Reznitskaya, A., Kuo, L., Glina, M., & Anderson, R. C. (2009). Measuring argumentative reasoning: What's behind the numbers?. *Learning and Individual Differences, 19*(2), 219–224. doi:10.1016/j.lindif.2008.11.001.

Şanlı, E. (2010). *Comparing reliability levels of scoring of the holistic and analytic rubrics in evaluating the scientific process skills*. (Unpublished master thesis). Ankara University, Ankara.

Şahin, M. G. (2019). Performansa dayalı değerlendirme [Performance-based assessment]. B. Çetin (Ed.). In *Eğitimde ölçme ve değerlendirme [Measurement and assessment in education]* (p. 213-264). Ankara: Anı Publ.

Şahin, M. G., & Boztunç-Öztürk, N. (2018). Scale development process in educational field: A content analysis research. *Kastamonu Education Journal, 26*(1), 191-199. doi: 10.24106/kefdergi.375863

Tindal, G., & Haladyna, T. M. (2012*). Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, New Jersey: Lawrence Erlbaum.

Turgut, F. (1992). *Eğitimde ölçme ve değerlendirme [Measurement and assessment in education]* (8. ed.). Ankara: Saydam Publ.

Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri [Qulatitative research methods in social sciences]* (9. ed.). Ankara: Seçkin Publ.

Yıldıztekin, B. (2014). *The comparison of interrater reliability by using estimating tecniques in classical test theory and generalizability theory.* (Unpublished master thesis). Hacettepe University, Ankara.

# Appendices

## Appendices 1. Articles Included in the Research

| ***Surname, N. (Year)*** | ***Research Name*** | ***Volume Number*** | ***Page Number*** |
|---|---|---|---|
| ***Journal A*** | | | |
| Alkaş Ulusoy, C. (2020) | Effects of number sense-based instruction on sixth-grade students' self-efficacy and performance. | *45*(202) | 417-439 |
| Aşık, G., & Erktin E. (2019) | Metacognitive experiences: mediating the relationship between metacognitive knowledge and problem solving. | *44*(197) | 85-103 |
| Altun, M., & Bozkurt, I. (2017) | A new classification proposal for mathematical literacy problems. | *42*(190) | 171-188 |
| Çetin, B., & İlhan, M. (2017) | An analysis of rater severity and leniency in open-ended mathematic questions rated through standard rubrics and rubrics based on the SOLO taxonomy. | *42*(189) | 217-247 |
| Çetinkaya, L. (2019) | The effects of problem based mathematics teaching through mobile applications on success. | *44*(197) | 65-84 |
| Romero Albaladejo, I. M., del Mar Garcia, M., & Codina, A. (2015) | Developing mathematical competencies in secondary students by introducing dynamic geometry systems in the classroom. | *40*(177) | 43-58 |
| Sarı, M. H., & Tertemiz, N. (2017) | The effects of using geometry activities based on Dienes' principles on 4th graders' success and retention of learning. | *42*(190) | 1-23 |
| ***Journal B*** | | | |
| Atalmış, E. H. (2018) | The use of three-option multiple choice items for classroom assessment. | *5*(2) | 314-324 |
| ***Journal C*** | | | |
| Altıntaş, E., & Özdemir, A. Ş. (2015) | The effect of developed differentiation approach on the achievements of the students. | 61 | 199-216 |
| Bal, A. P. (2016) | The effect of the differentiated teaching approach in the algebraic learning field on students' academic achievements. | 63 | 185-204 |
| Balta, E., & Ömür-Sünbül, S. (2017) | An investigation of ordering test items differently depending on their difficulty level by differential item functioning. | 72 | 23-42 |
| İlhan, A., Tutak, T., & Çelik, H. C. (2019) | What is the predictive power of visual mathematics literacy perception and its sub-dimensions for geometry success? | 80 | 1-24 |
| Özarkan, H. B., & Doğan, C. D. (2020) | A comparison of two standard-setting methods for tests consisting of constructed-response items. | 90 | 121-138 |
| Özdemir, A. Ş., & Sahal, M. (2018) | The effect of teaching integers through the problem posing approach on students' academic achievement and mathematics attitudes. | 78 | 117-138 |
| Özerem, A., & Akkoyunlu, B. (2015) | Learning environments designed according to learning styles and its effects on mathematics achievement. | 61 | 61-80 |
| Özmen, Z. M., Güven, B., & Kurak, Y. (2020) | Determining the graphical literacy levels of the 8th grade students. | 86 | 269-292 |
| Özyurt, H., & Özyurt, Ö. (2015) | Ability level estimation of students on probability unit via computerized adaptive testing. | 58 | 27-44 |
| ***Journal D*** | | | |
| Canbazoğlu, H. B., & Tarım, K. (2020) | An activity-based practice for improving mathematical literacy and awareness of elementary school teacher candidates. | *10*(4) | 1183-1218 |
| Elazzabi, A., & Kaçar, A. (2020) | Investigation of Libyan and Turkish students' thinking levels in solving quadratic word problems based on SOLO Taxonomy. | *10*(1) | 283-316 |

| | | | |
|---|---|---|---|
| Gür, H., & Hangül, T. (2015) | A study on secondary school students' problem solving strategies. | *5*(1) | 95-112 |
| İlhan, A., & Çelik, H. C. (2018) | Assessment of student achievement and views on the impact of instruction with visualization of identities in $(ax+b)^n$ form in mathematics. | *8*(4) | 833-878 |
| Kuzu, O. (2020) | Preservice mathematics teachers' competencies in the process of transformation between representations for the concept of limit: A qualitative study. | *10*(4) | 1037-1066 |
| Mecek, S., & Taşlıdere, E. (2015) | Investigation of gifted students mathematics and physics achievements in terms of different variables. | *5*(5) | 733-746 |
| Özcan, Z. C., & Doğan, H. (2018) | A longitudinal study of early math skills, reading comprehension and mathematical problem solving. | *8*(1) | 01-18 |
| Özdoğan, D., & Doğan, N. (2018) | The effect of the correction of self-assessment-based chance success on psychometric characteristics of the test. | *8*(3) | 567-598 |
| Tavşanlı, O. F., Kozaklı, T., & Kaldırım, A. (2018) | The effect of graphic organizers on the problem posing skills of 3rd grade elementary school students. | *8*(2) | 377–406 |
| Uyar, G., & Bal, A.P. (2015) | The effect of the problem based learning on academic success for the 6th grade students. | *5*(4) | 361-374 |
| ***Journal E*** | | | |
| Altun, A., & Kelecioğlu, H. (2016) | A comparison of calibration methods and proficiency estimators based on item response theory in vertical scaling. | *31*(3) | 447-460 |
| Deniz, Ö., & Uygur-Kabael, T. (2017) | Students' mathematization process of the concept of slope within the realistic mathematics education. | *32*(1) | 123-142. |
| Erkoç, A., & Dinç-Artut, P. (2016) | The effect of the team-assisted individualization technique on eighth grade students' geometry achievement and retention. | *31*(1) | 1-13 |
| Güner, P., & Uygun, T. (2020) | Examining students' mathematical understanding of patterns by Pirie-Kieren model. | *35*(3) | 644-661 |
| Gürbüz, R., Dede, Y., & Doğan, M. F. (2018) | The role of computer-assisted instruction in the teaching of probability. | *33*(3) | 705-722 |
| Gürefe, N. (2018) | Determining strategies used in area measurement problems by middle school students. | *33*(2) | 417-438 |
| İlhan, M. (2016). | A comparison of the ability estimations of classical test theory and the many facet Rasch model in measurements with open-ended questions. | *31*(2) | 346-368 |
| Karaaslan, G., & Turanlı, N. (2018). | An alternative tool for diagnosis of the misconceptions and the errors in complex numbers: Diagnostic branched tree method. | *33*(1) | 72-89 |
| Sevgi, S., & Cağlıköse, M. (2020). | Analyzing sixth-grade students' metacognition skills in process of solving fraction problems. | *35*(3) | 662-687 |
| Sevimli, E., & Delice, A. (2016). | Investigation of the influence of computer algebra system supported teaching from procedural-conceptual competencies: The case of integrals. | *31*(2) | 249-266 |
| Yavuz Mumcu, H., & Aktürk, T. (2020). | Mathematics teachers' understanding of the concept of radian. | *35*(2) | 320-337 |
| Yıldırım Yakar, Z., & Albayrak, M. (2018). | The effect of the layered curriculum method on the students' achievement in "area measurement". | *34*(2) | 565-585 |