Journal of Teacher Education and Lifelong Learning (TELL)

Volume: 5 Issue:2 Year: 2023 Research Article ISSN: 2687-5713

Developing an Achievement Test for Primary School English Course: Validity and Reliability Study

Işıl Akçay [1] **Nezih Önal** [2]

[1] Nigde Omer Halisdemir University, Department of Computer Education and Instructional Technology, Nigde, Turkey
isilakcay3638@gmail.com

[2] Nigde Omer Halisdemir University, Faculty of Education, Department of Computer Education and Instructional Technology, Nigde, Turkey
nezihonal@ohu.edu.tr

Article Info

Article History
Received: 12/08/2023
Accepted: 27/11/2023
Published: 27/11/2023

Keywords:
Achievement test development,
English language teaching,
Validity,
Reliability,
Item analysis.

ABSTRACT

The objective of this research was to develop a multiple-choice achievement test for the 4th-grade English lesson unit called "My Day." The test's administration took place in the autumn term of the 2022-2023 academic year and involved 621 fifth-grade students from the central district of Niğde province. Out of these students, 209 participated in the pilot phase, while the remaining 412 students were involved in the actual application of the test. The test items were meticulously analyzed using the Test Analysis Program (TAP). Based on the results of the item discrimination analysis, two items were deemed less effective and, therefore, excluded from the final version of the test. Ultimately, the test consisted of 25 carefully selected items to evaluate the student's understanding of the "My Day" unit. To assess the internal reliability of the test, the researchers calculated the KR-20 value which turned out to be 0.888. This value demonstrated that the test exhibited a satisfactory level of consistency and reliability in measuring the intended learning outcomes. Overall, the rigorous development process and statistical analysis provide confidence in the validity and accuracy of the achievement test. Several recommendations are made in order to increase the test's usefulness and efficacy. The researcher could conduct a follow-up study to evaluate the long-term impact of the achievement test on students' language learning progress. Additionally, they may explore adapting the test for use in various regions and cultures to assess its cross-cultural validity, aiming to further improve its effectiveness and applicability..

Citation: Akçay, I. & Önal, N. (2023). Developing an Achievement Test for Primary School English Course: Validity and Reliability Study. *Journal of Teacher Education and Lifelong Learning*, *5*(2), 778-788.

INTRODUCTION

With globalisation, commercial and cultural relations between countries have increased. Increasing relations have created a need for language learning, and the necessity of living in a multilingual world has become more understood. Today, English is one of the most widely used and learned languages among the world languages (Ilyosovna, 2020). English, which is accepted as a worldwide language, is used in many fields such as business, education and tourism. Knowing English increases competitiveness in the global market, expands international business opportunities and facilitates intercultural communication. In addition, many scientific articles and publications are written in English.

As in the whole world, the desire to learn English in Turkey continues to increase day by day. English language learning has become one of the indispensable elements of Turkish education policy (Sönmez & Köksal, 2022). The English language learning policy in Turkey emphasises the teaching of English from the primary school level onwards, and envisages its use as a primary language (Gürsoy, Korkmaz & Damar, 2017). In particular, learning English at an early age aims to catch the period when the child's language learning capacity is at its highest and to provide more permanent learning. In addition, an early start to language learning improves the child's linguistic skills and helps the formation of a wider language pool (Krasnıqı & Muhaxherı, 2019). For these reasons, studies on early language teaching in Turkey, as in EU countries, have increased in recent years. In 1997, with the extension of primary education to eight years, English was included among the compulsory courses in the fourth and fifth grades (Sarıçoban, 2012). With the uninterrupted 12-year compulsory education that started to be implemented as of the 2012-2013 academic year, English language teaching was included in the education programme starting from the second grade. In this way, it is aimed that children are introduced to a foreign language at an early age, gain awareness of language and culture, and begin to develop positive attitudes (Paker, 2018). The introduction of English education at an early age necessitated studies to meet the needs of young learners and some changes were made in the curriculum (Özüdoğru & Adıgüzel, 2015). With these changes, various methods, different approaches and applications have been used in English language teaching. In order to get efficiency from all these processes and to ensure that they can play an effective role in the lives of individuals, there should be a very good planning, implementation and measurement and evaluation system in English language teaching (Baysal & Ocak, 2019).

Assessment and evaluation in English language teaching is a process of determining how good students' English language skills are and in which areas they have improved. Assessment helps to see the effectiveness of English language teaching and the level of students' English language skills. Once the level of students' English language skills is determined, teachers can start to design a programme that focuses on English language teaching and is appropriate to students' needs. From this point of view, it is very important that assessment and evaluation should be carried out with quality and care (Meidasari, 2015). The detection and elimination of any teaching deficiency is only possible through a successful assessment and evaluation process. By using assessment and evaluation tools, teachers can determine how well students understand and how well they can apply. It also helps them to identify students' deficiencies and take the necessary steps to improve their areas of weakness. These are necessary steps for education to be efficient and effective. Measurement in education is the determination of the accuracy or level of a student's knowledge, skills and abilities (Kim, Raza, & Seidman, 2019). Assessment is the process of evaluating the knowledge and skills learned by a student according to certain criteria and making a judgment (Andrade & Brookhart, 2020). Measurement and evaluation in education are necessary to follow the learning process, identify deficiencies and strengths, encourage students and evaluate the quality of education. Measurement can be done through tests, exams, performance assessments, projects and other tools.

One of the most frequently used measurement tools is achievement tests. These tests are used to measure the knowledge and skills gained by students in a particular subject (Borghans & et al., 2016). The content of achievement tests is prepared depending on the course outcomes and administered at the end of the unit, semester or course (Alderson, Clapham, & Wall 2002). Achievement tests have higher validity and reliability than other measurement tools and this is one of the reasons why achievement tests are frequently preferred in education (Karip, 2012). English language teaching and assessment and evaluation are two disciplines that are very close to each other and achievement tests are a frequently used measurement tool in English language courses as in other courses. The validity and reliability of the measurement result depend on how valid and reliable the achievement test is. In this context, there is a need for achievement tests with proven validity and reliability in English language teaching as in every field. When the literature is reviewed, it is seen that there are very few studies in the field of English (Baysal & Ocak, 2019; Özüdoğru & Adıgüzel, 2015; İncirci & Parmaksız, 2016). In this context, the aim of this study is to develop an achievement test compatible with the objectives of the "My Day" unit in Grade 4 English lesson. The "My Day" unit aims to help fourth-grade students develop their organizational skills, time management, and self-awareness by planning and reflecting on their daily activities. It encourages students to understand the concept of routines and how they can be beneficial in managing their time and responsibilities effectively.

METHOD

Research Design

Developing an achievement test typically involves a research method known as test development or test construction. This method follows a systematic process to design, create, and validate a reliable and valid assessment tool to measure specific knowledge, skills, or abilities of the test takers. This study was conducted using quantitative research methods. (Fraenkel & Wallen, 2006).

Research Sample

The population of the study consisted of fifth-grade middle school students studying in the central district of Niğde, and the sample consisted of a total of 621 fifth-grade students (209 pilot, 412 actual implementations) studying in four middle schools in the central district of Niğde in the first semester of the 2022-2023 academic year. The gender distribution of the students is given in Table 1.

Table 1. *Gender information of the participant students*

Gender	n	%
Female	325	52
Male	296	48
Total	621	100

Although the test developed in the study was aimed at the fourth grade "My Day" unit, the reason why fifth-grade students were included in the sample of the study was that the fourth-grade students had not yet studied this unit, so it was thought that they would tend to leave the questions blank when a test was applied on a subject they did not know. In order to minimise this problem and to ensure that all questions were answered, the sample was composed of fifth-grade students.

Research Instruments and Processes

Determining the Purpose:

As a result of the literature review, no valid and reliable achievement test for the fourth-grade primary school English course was found. This achievement test was developed in order to measure student achievement towards the objectives of the fourth grade "My Day" unit. This study, which is original in the literature, aims to contribute to the literature and provide a data collection tool for future researchers.

Determination of the Learning Outcomes and Preparation of the Specification Table:

Before the achievement test questions were prepared, Bloom's taxonomy (Huitt, 2011) was taken into consideration and a specification table containing the learning outcomes was prepared. Since there were no open-ended questions in the achievement test aimed to be developed, it was not possible to prepare questions for application, analysis and synthesis steps.

Formation of the Question Pool:

A question pool consisting of 35 items was created by the researcher by taking into account the prepared specification table and the learning outcomes. Care was taken to prepare all the questions to be included in the achievement test in a way to represent all achievements of the target subject at a certain level. The learning outcomes of the questions and their distribution according to Bloom's taxonomy are given in Table 2.

Table 2. *Question pool specification table*

Subjects	Outcomes	Knowledge	Comprehension
1. Talking about the daily routine	1.1. Students will be able to understand the general and specific information in a short, oral text about daily routines.	1,2,13,28,31	16,19,27
	1.2. Students will be able to talk about their daily routines.	4,12,22,29,33	8,11,25,30
2. Telling the time and days	2.1. Students will be able to recognize the time in a short oral text.	3,5,17,21,24,35	9,20,23
	2.2. Students will be able to talk about the time.	6,14,15,18,26,34	7,10,32

Obtaining Expert Opinions, Writing Supervision and Revision of the Items:

Expert opinions were sought to ensure the content, construct and face validity of the prepared questions. The opinions of three teachers who are experts in the field of English were taken for the content validity study, an associate professor in the field of measurement and evaluation and an associate professor in the field of instructional technologies for the construct validity study, and a teacher who is an expert in the field of Turkish Language and Literature for the face validity study. As a result of the interviews, the items were reviewed one by one and the number of items was reduced to 27 because some of the questions did not measure the outcomes, the visuals could not be understood, and 35 questions were unlikely to be completed by primary school students in one lesson hour. The final version of the specification table after the expert opinion is given in Table 3.

Table 3. *Specification table after expert opinion*

Subjects	Outcomes	Knowledge	Comprehension
1. Talking about the daily routine	1.1. Students will be able to understand the general and specific information in a short, oral text about daily routines.	1,10,22,23	14,21
	1.2. Students will be able to talk about their daily routines.	3,9,16,25	6,19
2. Telling the time and days	2.1. Students will be able to recognize the time in a short oral text.	2,12,15,18,27	7,17
	2.2. Students will be able to talk about the time.	4,11,13,20, 26	5,8,24

Pilot Application and Item Analysis:

The achievement test, which was reduced to 27 items after the expert opinion, was applied in printed paper form to 209 fifth-grade students studying in a secondary school in the central district of Niğde in the first semester of the 2022-2023 academic year with the permission of the Directorate of National Education. The sample was selected by random sampling method. Students were given 1 lesson hour for the test and it was determined that it was completed in the given time. After the pilot application, the data were entered into the Excel matrix prepared by the researcher and analysed with the TAP programme. Item difficulty and discrimination indices of each item were analysed.

In test analysis, the item difficulty index (P) is an indicator that measures how difficult a question in a test is. It is usually calculated as the correct answer rate of the question and low rates indicate that the question is more difficult and high rates indicate that it is easier (Tekin, 2000). Evaluation criteria according to the item difficulty index are given in Table 4.

Table 4. *Evaluation criteria according to item difficulty index*

Item Difficulty Index (P)	Item Evaluation
0.00-0.29	Difficult
0.30-0.49	Medium difficulty
0.50-0.69	Easy
0.70-1.00	Very easy

Item difficulty index (P) takes a value between 0 and 1. When the P value approaches zero, it indicates that the item is difficult and when it approaches one, it indicates that the item is easy. It is desirable that the item difficulty index of the item is around 0.50, that is, the item should be neither too difficult nor too easy.

In test analysis, item discrimination index (D) is an indicator that measures how a question in a test affects the performance of a group of students. It is usually calculated as the difference between the correct answer rate of students with better performance and the correct answer rate of students with worse performance (Büyüköztürk, 2011). Evaluation criteria according to item discrimination index are given in Table 5.

Table 5. *Evaluation criteria according to item discrimination index*

Item Distinctiveness Index (D)	Item Selection decision
0.19 and smaller	Should not be tested or replaced completely
Between 0.20-0.29	Should be corrected and tested
Between 0.30-0.39	Should be tested without correction or with minor adjustments
0.40 and greater	Good item should be tested as is

Item discrimination index (D) takes a value between -1 and +1. When the D value approaches -1, it indicates that the discrimination is low, and when it approaches +1, it indicates that the discrimination is high. The higher the discrimination, the higher the reliability of the item.

Conducting the Actual Application and Item Analysis:

In the item analysis conducted after the pilot application, two questions were removed from the test because the discrimination index of two questions was below 0.20. Without making any changes in the other questions, the achievement test was made ready for the actual application with 25 multiple-choice questions. The actual application of achievement test was applied to a total of 412 fifth-grade students studying in three secondary schools in the central district of Niğde in the first semester of the 2022-2023 academic year. After the actual application, the data were entered into the Excel matrix prepared by the researcher and analysed with the TAP programme. The item difficulty and discrimination indices of each item were analysed. The scores of the 25-item achievement test applied

to the students and the calculated test statistics are given in Table 6.

Table 6. *Test statistics*

	Score
Number of Examinees	412
Total Possible Score	25
Minimum Score	1.000 = 4.0 %
Maximum Score	25.000 = 100 %
Median Score	17.000 = 68.0%
Mean Score	16.733 = 66.9%
Standard Deviation	5.767
Variance	33.264
Skewness	0.418
Kurtosis	-0.847
Mean Item Difficulty	0.669
Mean Discrimination Index	0.550
Mean Point Biserial	0.524
KR20 (Alpha)	0.868

Ethic

The author(s) confirm(s) that ethical approval was obtained from Niğde Ömer Halisdemir University (Approval Date: 26 /10 /2022, 2022/12-28).

FINDINGS / RESULTS

Findings Related to Item Analysis

The results of the item analyses of the pilot application of the achievement test are given in Table 7 and the results of the item analyses for the actual application are given in Table 8.

Table 7. *Pilot application item analysis results*

Item	**Number correct**	**Item Diff.**	**Disc. Index**	**Correct in high grp**	**Correct in low grp**	**Point biserial**
1	126	0.60	0.55	55	20	0.49
2	141	0.67	0.48	54	23	0.42
3	109	0.52	0.48	49	18	0.37
4	71	0.34	0.20	30	17	0.16
5	144	0.69	0.67	63	20	0.61
6	70	0.33	0.39	35	10	0.39
7	88	0.42	0.52	44	11	0.43
8	107	0.51	0.58	52	15	0.47
9	73	0.35	0.55	44	9	0.51
10	158	0.76	0.50	62	30	0.45
11	128	0.61	0.59	56	18	0.54
12	85	0.41	0.62	46	7	0.53
13	157	0.75	0.58	61	24	0.54
14	102	0.49	0.69	56	12	0.57
15	46	0.22	0.14	22	13	0.19
16	154	0.74	0.61	63	24	0.54
17	64	0.31	0.39	36	11	0.40
18	116	0.56	0.58	55	18	0.49
19	121	0.58	0.63	58	18	0.54
20	96	0.46	0.70	54	9	0.56
21	110	0.53	0.70	56	11	0.61

22	107	0.51	0.73	58	11	0.61
23	91	0.44	0.52	49	16	0.48
24	85	0.41	0.47	43	13	0.46
25	128	0.61	0.66	56	14	0.54
26	77	0.37	0.47	42	12	0.43
27	78	0.37	0.61	46	7	0.54

According to the item analysis results of the pilot application consisting of 27 items, the standard deviation value of the test was calculated as 6,174. The standard deviation value of an achievement test shows how variable the test scores are (Tabachnick, Fidell, & Ullman, 2007). A high standard deviation indicates that the distribution of test results is wide and that students perform differently on the test. In this case, the standard deviation value of 6.174 showed that the test results varied slightly among the students and the questions in the test were easily answered by some students, while some students had difficulty.

The Point biserial correlation coefficient measures the relationship between the correct answer of a question and the overall test scores and takes a value between -1 and 1. The higher the point biserial value of a question, the higher the correlation between the correct answers and the overall test scores. If the point biserial value is close to zero, it means that there is no relationship between the correct answer of the question and the overall test scores (Kornbrot, 2014). The point biserial value of the achievement test was calculated as 0.476 and showed that the test was a medium level test.

The average difficulty index of the achievement test was calculated as 0,502. Item difficulty index is a parameter that measures how difficult or easy a test is. Item difficulty index takes a value between 0 and 1 and 0.5 represents a moderate level of difficulty (Tekin, 2000). Therefore, the item difficulty index of 0.502 indicated that this achievement test had an average level of difficulty.

The average discrimination index of the achievement test was calculated as 0,542. Item discrimination index is a parameter that measures whether a test can distinguish students with high scores from students with low scores. The item discrimination index takes a value between 0 and 1 and values higher than 0.40 are considered as good discrimination (Büyüköztürk, 2011). Therefore, the item discrimination index of 0.542 showed that this achievement test had a moderate level of discrimination.

After the pilot application, item analysis and option analyses were performed with TAP software and discrimination index and point biserial values of each item were examined. The lowest item discrimination index was 0,141 (M14) and the highest discrimination index was 0,734 (M16). The lowest point biserial value was 0,164 (M4) and the highest point biserial value was 0,611 (M21). According to the evaluations of item discrimination and point biserial indices, items with a value less than 0,20 should either be removed from the test or changed completely. Accordingly, M4 and M14 were submitted to expert opinions again and it was decided to remove the items from the test. Before the items were removed, the specification table was examined and it was determined that there was no problem in terms of content validity. The number of items was rearranged and the 25-item achievement test was made ready for the actual application.

Table 8. *Main application item analysis results*

Item	**Number correct**	**Item Diff.**	**Disc. Index**	**Correct in high grp**	**Correct in low grp**	**Point biserial**
1	344	0.83	0.44	127	67	0.51
2	362	0.88	0.33	127	81	0.44
3	344	0.83	0.46	127	65	0.51
4	337	0.82	0.53	128	57	0.58
5	181	0.44	0.63	99	17	0.52
6	220	0.53	0.50	102	36	0.42
7	210	0.51	0.56	103	30	0.48
8	220	0.53	0.66	116	30	0.51
9	349	0.85	0.43	128	69	0.54

10	253	0.61	0.63	118	36	0.56
11	231	0.56	0.65	110	26	0.53
12	364	0.88	0.31	127	83	0.48
13	268	0.65	0.77	127	27	0.65
14	370	0.90	0.26	127	89	0.40
15	178	0.43	0.55	127	89	0.45
16	261	0.63	0.61	123	43	0.53
17	268	0.65	0.62	123	41	0.53
18	289	0.70	0.50	120	53	0.48
19	308	0.75	0.67	127	39	0.65
20	328	0.80	0.58	127	50	0.61
21	243	0.59	0.76	124	26	0.63
22	194	0.47	0.66	109	23	0.53
23	352	0.85	0.44	128	68	0.55
24	165	0.40	0.47	89	28	0.36
25	255	0.62	0.72	124	30	0.64

According to the item analysis results of the actual application consisting of 25 items, the standard deviation value of the test was calculated as 5,767. The lower standard deviation value compared to the pilot application showed that the test results varied less among the students.

The point biserial value and discrimination index of the final achievement test were calculated as 0,524 and 0,550, respectively. The higher values compared to the pilot application showed that the discrimination of the final achievement test with 25 items was higher. The average difficulty index of the final achievement test was calculated as 0,669. The higher value compared to the pilot study showed that the achievement test was easier. The histogram graph of the students' 25-item achievement test scores is given in Figure 1.

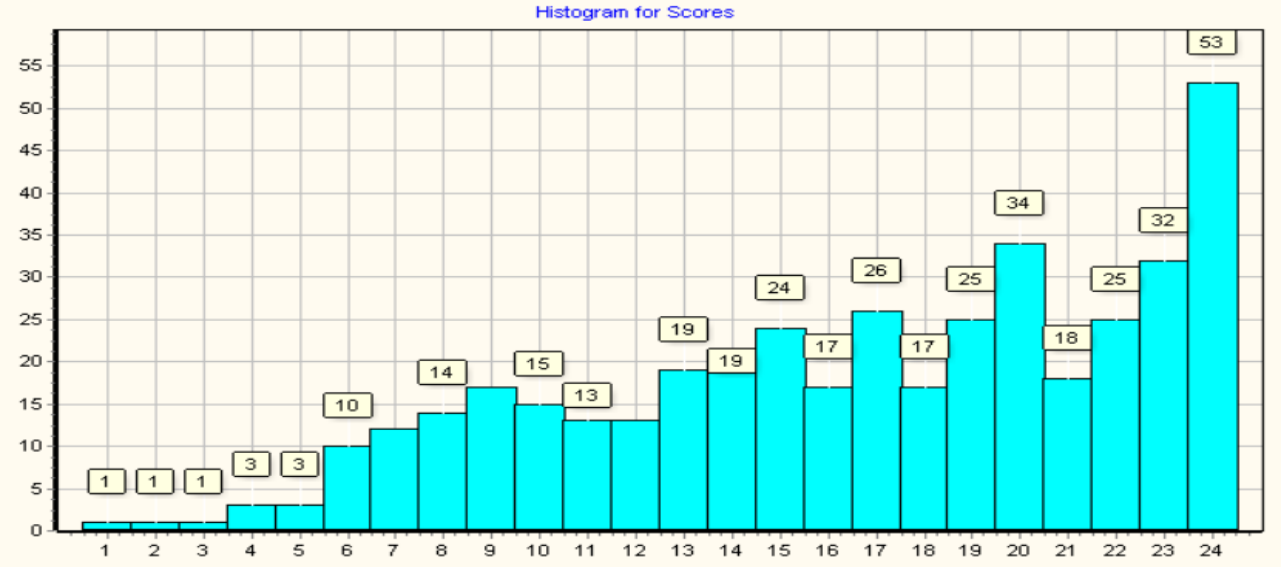

Figure 1. *Histogram graph based on assessment scores*

In the achievement test, kurtosis and skewness values are statistical measures that measure the shape and symmetry of the distribution of test scores. If the kurtosis and skewness values of the test scores are between +1.5 and -1.5, a normal distribution is observed (Tabachnick, Fidell & Ullman, 2007). As a result of the application, the kurtosis value of the achievement test was calculated as -0.847 and the skewness value as 0.418. In this context, it was observed that the achievement test scores were normally distributed.

After the actual application, item analysis and option analyses were performed with the TAP programme, and the discrimination index and point biserial values of each item were examined. The lowest item discrimination index was 0,263 (M14) and the highest discrimination index was 0,771 (M13). The lowest point biserial value was 0,364 (M14) and the highest point biserial value was 0,651

(M13). Considering that the items with item discrimination and point biserial values higher than 0.30 have good discrimination (Usta and Karakuş, 2016), no changes were made to the items.

Findings Related to Item Analysis

Reliability in achievement tests is a feature that determines the repeatability of the test and how accurate and stable its measurement is (Mohamad & et al., 2015). In other words, the reliability of an achievement test indicates the consistency of the results of the same test when it is applied at different times or under different conditions. Kuder-Richardson 20 formula was used to measure the reliability of the achievement test.

KR-20 Reliability test is a method used to measure intra-test consistency. This test is particularly suitable for multiple-choice tests. The KR-20 value is a number ranging from 0 to 1, and the closer it is to 1, the higher reliability the test is considered to have. If the KR-20 value is 0.70 and above, the test is considered to have sufficient reliability (Kılıç, 2016).

The KR-20 value of the pilot application was calculated as 0.870; the KR-20 value of the actual application was calculated as 0.888. In this context, it was observed that both application tests were quite reliable, but the actual application was more reliable.

DISCUSSION, CONCLUSION, RECOMMENDATIONS

Measurement and evaluation in foreign language teaching is an important tool that helps to determine where students are in the learning process, to guide the teaching process for teachers, to determine students' needs and to increase their motivation. Achievement tests are one of the most frequently used methods in measurement and evaluation. It is important that achievement tests are valid and reliable in order to accurately measure students' actual achievement. A test that is not valid and reliable can prevent educators from making the right decisions by mismeasuring students' achievement. In the literature review, it was found that there are very few achievement tests with proven validity and reliability in the field of English (Baysal & Ocak, 2019; Özüdoğru & Adıgüzel, 2015; İncirci, 2016). This study, it was aimed to develop an assessment tool compatible with the 4th grade English lesson "My Day" unit outcomes and to reveal the item analyses of this tool.

Considering the test development steps, firstly the aims of the test were determined and the target group was selected. A specification table suitable for the achievements of the unit for which the achievement test was to be prepared was prepared and a question pool of 35 items was created. After the expert opinions, the number of items was reduced to 27 and a pilot application was conducted with 209 5th grade students. The KR-20 value of the pilot application was 0.80 and the average discrimination index was calculated as 0.542. According to the results of the item analysis performed in the TAP programme, since the discrimination index of 2 items was below 0.20, the items were presented to the expert opinion again and since it was not possible to make changes in the items, it was decided to remove the items from the test. The achievement test, which was reduced to 25 items, was applied to 412 fifth-grade students. The KR-20 value of the actual application was 0,888 and the average discrimination index was calculated as 0,550. According to the results of the item discrimination index, all of the items were higher than 0,30 and there was no need to make any changes.

In this study, a multiple-choice achievement test with proven validity and reliability was developed. Since there is no valid and reliable test development study for the 4th grade English course in the literature, it is thought that the achievement test will contribute to this field.

As a result, this achievement test is important both in terms of providing a valid and reliable measurement tool for primary school English teachers in the process of evaluating their students and in terms of providing a data collection tool for lecturers and researchers conducting scientific studies in the

field of language. In order to standardise the test, the number of samples to which the test will be applied can be increased or the method of applying the test can be changed. It is recommended for future researchers to develop achievement tests for other unit achievements and to conduct a follow-up study to assess the long-term impact of the achievement test on students' language learning progress. Researchers may explore the possibility of adapting the test for use in different regions and cultures to evaluate its cross-cultural validity or investigate the potential benefits of incorporating different question formats, such as open-ended or performance-based items, to assess students' English language proficiency comprehensively.

Limitations

This study on the development of an achievement test for the fourth-grade English lesson "My Day" unit presents several limitations. Firstly, the sample size was limited to 621 fifth-grade students from the central district of Niğde province, which may restrict the generalizability of the findings to a broader population. Additionally, the study focused on a specific grade level, overlooking potential variations in language learning progress across different educational stages. The study did not account for external factors like student motivation or socio-economic backgrounds, which could influence test performance

REFERENCES

Alderson, C., Clapham, C., & Wall, D. (2002). *Language test construction and evaluation* (9th ed.). Cambridge University Press.

Andrade, H. L., & Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assessment In Education: Principles, Policy & Practice, 27*(4), 350-372. https://doi.org/10.1080/0969594X.2019.1571992

Baysal, A. E., & Ocak, G. (2019). An achievement test on “studying abroad” and “my environment” units: the study of validity and reliability. *International Journal of Social Science Research, 8*(1), 1-19. https://dergipark.org.tr/en/pub/ijssresearch/issue/46587/402223

Borghans, L., Golsteyn, B. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure? *Proceedings of the National Academy of Sciences, 113*(47). https://doi.org/10.1073/pnas.1601135113

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

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). McGraw-Hill.

Gursoy, E., Korkmaz, S. C., & Damar, E. A. (2017). English language teaching within the new educational policy of Turkey: Views of Stakeholders. *International Education Studies*, *10*(4), 18-30. https://doi.org/10.5539/ies.v10n4p18

Huitt, W. (2011). Bloom et al.'s taxonomy of the cognitive domain. *Educational Psychology Interactive, 22,* 1-4. http://www.edpsycinteractive.org/topics/cognition/bloom.pdf

Ilyosovna, N. A. (2020). The importance of English language. *International Journal on Orange Technologies, 2*(1),22-24. https://researchparks.innovativeacademicjournals.com/index.php/IJOT/article/view/4730

İncirci, A., & Parmaksız, R.Ş. (2016). Development of achievement test related to English class “simple past tense achievement test”. *International Journal of Language Academy*, *4*(2). https://doi.org/10.18033/ijla.408

Karip, E. (2012). *Ölçme ve değerlendirme*. Pegem A Yayıncılık.

Kılıç, S. (2016). Cronbach's alpha reliability coefficient. *Psychiatry and Behavioral Sciences*, *6*(1),47 https://doi.org/10.5455/jmood.20160307122823

Kim, S., Raza, M., & Seidman, E. (2019). Improving 21st-century teaching skills: The key to effective 21st-century learners. *Research in Comparative and International Education*, *14*(1), 99-117. https://doi.org/10.1177/1745499919829214

Kornbrot, D. (2014). Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat06227

Krasnıqı, S., & Muhaxherı, N. (2019). Issues and benefits of the literature inclusion into English language teaching. *Journal of International Social Research*, *12*(68). http://dx.doi.org/10.17719/jisr.2019.3823

Meidasari, V. E. (2015). The assessment and evaluation in teaching English as a foreign language. *Indonesian EFL Journal*, *1*(2), 224-231. https://doi.org/10.25134/ieflj.v1i2.629

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia-Social and Behavioral Sciences*, *204*, 164-171. https://doi.org/10.1016/j.sbspro.2015.08.129

Özüdoğru, F., & Adıgüzel, O. C. (2015). Development of a listening and speaking achievement test for primary school 2nd grade English course. *Pegem Education and Training Journal, 5*(4), 375-396. http://dx.doi.org/10.14527/pegegog.2015.021.

Paker, T. (2018). İngilizce öğretiminde ölçme ve değerlendirme. E. Üstünel & Ş. Kömür (eds.), *Kuramdan uygulamaya sınıf öğretmenliği seti: ilkokulda yabancı dil öğretimi,* (p145-153).Eğiten Kitap.

Saricoban, G. (2012). Foreign language education policies in Turkey. *Procedia-Social and Behavioral Sciences, 46*, 2643-2648. https://doi.org/10.1016/j.sbspro.2012.05.539

Sönmez, G., & Köksal, O. (2022). A critical overview of English language education policy in Turkey. *Shanlax International Journal of Education*, *11*(1), 1-9. https://doi.org/10.34293/ education.v11i1.5721

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics.*(Vol. 5). Pearson.

Tekin, H. (2000). *Eğitimde ölçme ve değerlendirme.* Yargı Yayınları.

Usta, M. E., & Karakuş, M. (2016). The development of the scale of pedagogical literacy. *Kastamonu Education Journal, 24* (1) , 133-14. https://dergipark.org.tr/tr/pub/kefdergi/issue/22606/241622