

Development of a Three-Tier Diagnostic Test on Simple Machines for Pre-Service Science Teachers¹

Damla Erken², Nazan Ocak İskeleli³, Cumhuri Türk⁴

Abstract

The aim of this study was to develop a valid and reliable three-tier diagnostic test to identify misconceptions about simple machines among science teacher candidates in the Physics-1 course. The study group consisted of 126 teacher candidates enrolled in the science education program at a state university in the Black Sea region during the 2024–25 academic year. The researchers developed a data collection tool called the "Three-Tier Diagnostic Test on Simple Machines." All tiers of the test consisted of multiple-choice questions. During test development, the purpose of the test was first determined, and then important behaviors were included in the scope of the test to create an indicator table. Twenty-one trial test items were prepared according to the revised Bloom Taxonomy levels. The test was submitted for expert review, and after making the necessary revisions, the pilot application of the three-tier diagnostic test was conducted. The data obtained from the pilot study were analyzed using the TestAn software package. As a result of the analysis, a valid and reliable 17-item "Simple Machines Three-Tier Diagnostic Test" was developed. The three-tier diagnostic test's reliability coefficient (KR-20) was found to be 0.71. In other words, the reliability of the test is considered reliable when it is above 0.70. It is recommended that researchers expand the three-tier diagnostic test for science teacher candidates to a four-tier test, test its reliability by applying it to different working groups, or create a three-tier diagnostic test on simple machines for middle and high school students.

Received:

01 October 2025

Accepted:

15 December 2025

Published online:

19 December 2025

Keywords

Science education,
Simple machines,
Science teacher
candidates,
Misconceptions,
three-tier diagnostic test

To cite this article: Erken, D., Ocak İskeleli, N., & Türk, C. (2025). Development of a three-tier diagnostic test on simple machines for pre-service science teachers. *Journal of STEM Teacher Institutes*, 5(2), 36-50. Retrieved from <https://jstei.com/index.php/jsti/article/view/95>

¹ This article was produced from the first author's master's thesis.

² Ondokuz Mayıs University, dmalerken@gmail.com, Orcid ID: 0009-0008-5628-5873

³ Ondokuz Mayıs University, nocak@omu.edu.tr, Orcid ID: 0000-0001-5794-3681

⁴ Samsun University, cumhur.turk@samsun.edu.tr, Orcid ID: 0000-0002-8630-9353

Introduction

Concepts play a critical role in science education. Used by individuals to make sense of events and their relationships, concepts have an important place in science and physics education (Ayvacic & Devvecoglu, 2009). Concepts are the abstract representations in our minds of specific groups that we create based on particular characteristics (Ayas et al., 1997). It is essential that concepts be taught concretely and in a clear and understandable way in educational activities. Science education involves many abstract, difficult-to-understand concepts generally related to physics topics (e.g., physical events) (Dağdalan & Taş, 2017; Aci-Özkan, 2024).

Abstract and difficult-to-understand concepts may form in students' minds in ways that differ from what is anticipated. Various studies show that students develop ideas and beliefs about certain concepts and phenomena before receiving formal science education and bring these ideas with them into their educational lives (Amir & Tamir, 1994). In other words, students may enter science education with ideas and beliefs based on individual experiences and simple observations that do not align with scientific facts. Such ideas are referred to as "conceptual misconceptions" in the literature (Laçin-Şimşek, 2019).

These misconceptions are considered an important issue to address in the learning and teaching process (Gülçiçek & Yağbasan, 2004). Eliminating misconceptions is important for students to understand the natural world and provide explanations for events, as well as for meaningful, lasting learning (Avcı et al., 2012). A review of the literature indicates that misconceptions exist among individuals of all ages (students, teacher candidates, and teachers) (Ayvacı et al., 2004; Bostan, 2008; Boz, 2005). Furthermore, studies have found that the same misconceptions exist in different age groups, negatively affecting lasting learning (Avcı et al., 2012). Identifying misconceptions in science education plays an important role in ensuring effective learning. A three-tier diagnostic test mainly consisting of multiple-choice questions can be used for this purpose (Çetinkaya & Taş, 2016; Özmen & Sever, 2024; Peker & Taş, 2019; Yıldırım & Semiz, 2022).

These tests reveal the reasons behind students' or teacher candidates' answers on any topic covered in class, i.e., they identify misconceptions (Peker & Taş, 2019). In the first tier, students are asked a question with four or five distractors, similar to multiple-choice questions. In the second tier, students explain why they selected a distractor in the first tier or mark one of several justifications consisting of four or five distractors. Additionally, a blank space is provided. Students are expected to use this space to explain any additional reasons they wish to add. In the third and final tier, students indicate their level of confidence in the answers they provided in the first two tiers (Aykutlu & Şen, 2012; Rudi et al., 2021).

If a student gives an incorrect answer in either of the first two tiers and is confident in their answer, they are said to have a "conceptual misconception." If a student answers incorrectly in either of the first two tiers but is not confident, they are said to have a "knowledge gap." If a student answers correctly in both tiers one and two and is confident, they are determined to have "scientific knowledge." However, if they are not confident, they are considered to have "given a lucky guess or lack confidence" (Peşman & Eryılmaz, 2010).

Bloom's taxonomy plays an important role in test development. Designed by Benjamin Bloom (1956), this taxonomy measures individuals' cognitive behaviors. These cognitive behaviors are measured using Bloom's six stages: knowledge, comprehension, application, analysis, synthesis, and evaluation (Colletta & Chiappetta, 1989; Krathwohl, 2002; cited in Soylu et al., 2020). In 1995, Lorin W. Anderson (a student of Bloom's) and his colleagues revised Bloom's taxonomy, creating a new one that differed from the original. One significant difference is that the six main levels were changed from nouns to verbs. Another important difference is that the revised taxonomy was reorganized from a one-dimensional structure to a two-dimensional structure consisting of a cognitive processes dimension and a knowledge type dimension. The cognitive process/area dimension includes mental processes ranging from recall to creation. The knowledge dimension consists of four categories: factual, conceptual, procedural, and metacognitive knowledge (Ersoy & Bayraktar, 2018). Bloom's Taxonomy is often

chosen by researchers because it is suitable for identifying students' or teacher candidates' cognitive process skills when the test's validity and reliability are met (Büyüköztürk, 2013).

Science education teaches a range of concrete and abstract concepts. A review of the literature reveals that students have conceptual misconceptions, particularly in studies related to simple machines. Researchers such as Avcı et al. (2012), İspir (2020), İspir and Aydın (2020), and Polat and Gödek (2022) have identified these misconceptions. This study used the cognitive knowledge levels of the "Revised Bloom Taxonomy" for the questions of a three-tier diagnostic test developed to identify misconceptions held by science teacher candidates.

"Simple machines" is a topic that students often struggle with and where misconceptions are prevalent. A review of the literature reveals that teacher candidates and students experience difficulty understanding simple machines, anxiety during the learning process, and struggle with teaching concepts. They also have low conceptual understanding and hold various misconceptions (Acar et al., 2019; Avcı et al., 2012; Bilgin & Kala, 2018; Diken, 2018). Therefore, it is essential to address this issue by developing suitable teaching materials and examining the impact of various teaching methods on cognitive skills.

Research Purpose and Significance

The purpose of this research is to develop a valid and reliable measurement tool that can identify science teacher candidates' misconceptions about simple machines, while taking the stages of test development into account. A review of the literature reveals that there are few studies on identifying misconceptions by administering a three-tier diagnostic test on simple machines to science teacher candidates. Therefore, this study is anticipated to fill a gap in the field.

Method

Research Design

This study aimed to develop a valid and reliable test consisting of three-tier, multiple-choice items to identify science teacher candidates' misconceptions about simple

machines, a topic covered in Physics 1. According to Baykul (2015), test development is the process of creating a test with predefined characteristics that provides measurable results according to a specific method.

The Simple Machines Three-Tier Diagnostic Test was developed based on the outcomes of the Science Teaching Undergraduate Program determined by the Council of Higher Education (YÖK). After creating the item pool, expert opinions were sought to ensure the content validity of the test. For example, the Turkish language expert provided the following feedback on the second question: "This question is generally understandable. There are no errors in the wording of the question or the answer choices. However, since 'usta' is used as a title for a person in question, it should begin with a capital letter according to spelling rules. A comma should be used instead of a semicolon in the answer choices in the 'Reason for choosing this answer' section." Similarly, a faculty member specializing in science education provided the following feedback on the second question: "The question is quite good in terms of the science field and corresponds to the given learning outcomes. However, it would be better if it were written as 'Friction on inclined planes has been neglected' in parentheses. Furthermore, this question is appropriate for the evaluation stage of Bloom's taxonomy. This is because, considering the length of the inclined plane, it involves making judgments and evaluations about force and force gain." After making the necessary adjustments based on the experts' opinions, the Simple Machines Three-Tier Diagnostic Test was piloted, and the data collected were subjected to item analysis. The item analysis established the content validity of the Simple Machines Three-Tier Diagnostic Test, and the reliability values were calculated using the KR-20 formula, yielding a value of 0.71. Based on the data analysis, items 6, 10, 14, and 21 were removed from the test, and the final version of the Simple Machines Three-Tier Diagnostic Test was created. Item 10, which had a discrimination and difficulty value of 0.12, is an example of an item removed from the study. According to Crocker and Algina (2006), item discrimination is considered low when it is below 0.30. Therefore, item 10 was removed.

Study Group

This research study consists of science teacher candidates at a state university in the Black Sea region during the 2024–25 academic year. A total of 126 teacher candidates have taken Physics-1. Due to considerations of economy and time efficiency (Erkuş, 2017; Yıldırım & Şimşek, 2021), the "convenience sampling" method was chosen for the research sample.

Data Collection Tool and Development Process

The data collection tool used in the study was the Simple Machines Three-Tier Diagnostic Test (BMÜATT), which the researcher developed to identify misconceptions held by science teacher candidates. The test consists of 21 multiple-choice questions developed with expert input.

The primary purpose of the three-tier diagnostic test is to reveal teacher candidates' misconceptions. The test development process consists of specific stages, and the test's validity and reliability increase as each stage is successfully completed. There are many comprehensive studies in the literature regarding these stages. In this study, the test development stages proposed by Atilgan et al. (2015, pp. 316–334) were considered. These stages are as follows:

- 1- Determining the purpose of using test scores

- 2- Selecting the achievements that correspond to the structure or area and preparing the indicator table
- 3- Writing the draft items
- 4- Reviewing (checking) the draft items
- 5- Creating the draft test form
- 6- Implementation of the draft test
- 7- Determining the items through item analysis procedures following the pilot application
- 8- Determining the statistical values of the final test created from the identified items.

The learning outcomes related to the topic of simple machines covered in Physics-1 are listed as follows:

- 1. The general characteristics of simple machines / the advantages they provide (no gain in work, providing ease of work in daily life) are emphasized.
- 2. Simple machines, such as fixed pulleys, movable pulleys, blocks and tackle, levers, inclined planes, and wheels and axles, are discussed.
- 3. Simple machines such as gears, screws, and pulleys are discussed.

The alignment of the learning outcomes for the topic of simple machines with the three-tier diagnostic test questions is presented in Table 1 below.

Table 1

Matching the Learning Outcomes of the Simple Machines Topic with Three-Step Diagnostic Test Questions

Question Number Related to the Learning Outcome	Question Number
1	1, 6, 14, 16
2	2, 5, 7, 9, 11, 12, 13, 15, 17, 18, 19, 20
3	3, 4, 8, 10, 21

The indicator table for the test, developed based on the Revised Bloom Taxonomy and

consisting of 21 pilot items, is presented in Table 2 below.

Table 2

Itemization Table of Test Items

Subject Name	Learning Outcomes	Cognitive Domain Levels					Total Number of Questions	
		Recall	Comprehension	Application	Analysis	Evaluation		
Simple Machines	1. The general characteristics of simple machines and the advantages they offer (no gain from work, providing convenience in daily life) are emphasized.	1	1		1	1	4	
	2. Simple machines, fixed pulleys, movable pulleys, blocks and tackle, levers, inclined planes, and wheels and axles are discussed.		3	3	2	3	1	12
	3. Simple machines, such as gears, screws, and pulleys, are discussed.	1		3	1			5
Test Item Number		1, 3	5, 13, 14, 17	4, 8, 9, 11, 15, 21	6, 7, 10, 20	2, 12, 16, 19	18	
Total		2	4	6	4	4	1	21

The draft three-tier diagnostic test, consisting of 21 items, was submitted for review to one research assistant and one faculty member specializing in the field. Additionally, the clarity and comprehensibility of the question texts in the pilot three-tier diagnostic test were presented for review to one Turkish language teacher who is a specialist with a degree in Turkish education. The necessary revisions were made based on the expert opinions received. Subsequently, a pilot application of the prepared three-tier diagnostic test was conducted with 25 randomly selected teacher candidates at the undergraduate level, without any time constraints. As a result of the pilot application, considering the average completion times of the teacher candidates, it

was decided that the test would be allocated 45 minutes.

The test items were diversified in terms of difficulty levels, instructions and warnings were added, and the font size was adjusted to suit the level of the teacher candidates before being placed on the test form. The three-tier diagnostic test prepared in the first tier was administered to 25 science teacher candidates. During the application process, care was taken to ensure that teacher candidates did not leave any items blank. The classes (license levels) and school numbers of the teacher candidates were printed on the test to help them focus on the three-tier diagnostic test.

The pilot test was administered to science teacher candidates.

The pilot test was then analyzed. After the three-tier diagnostic test was analyzed using the TestAn program, items were selected for the final test. When selecting items, item discrimination and difficulty levels must be taken into account. According to Ebel (1995),

there are certain criteria that must be used when selecting pilot items. These are presented in the tables below.

According to the criteria in Table 3, items that should not be included in the test, items requiring revision, and items that can be included directly can be evaluated using the item discrimination index..

Table 3
Item Selection Criteria Based on Item Discrimination Indices

Item Discrimination Index	Item Selection Criteria
0.19 and below	Should definitely not be included in the test or should be completely revised.
Between 0.20 and 0.29	These items are at the test's threshold and may be included in the test if necessary after correction.
0.30 to 0.39	These items can be included in the test without correction or with minor corrections.
0.40 and above	These are very well-performing items and can be included in the test as they are.

This study eliminated items based on the item discrimination index. Examples of eliminated items include item 10, with an index of less than 0.19, and items 6, 14, and 21, with indexes between 0.20 and 0.29. Table 9 in the Findings section provides a detailed list of the items included and excluded from the test.

The item difficulty index is the average of correct responses to an item scored as 1 or 0, reflecting both an average value and a percentage. It is calculated by dividing the number of participants who answered correctly by the total number of participants who took the test (Atilgan et al., 2015). The criteria for evaluating items based on the item difficulty index are presented in Table 4 below.

Table 4
Evaluation of Items Based on Item Difficulty Index

Item Difficulty Index	Item Evaluation
0.00 – 0.29	Difficult item
0.30 – 0.49	Moderately difficult item
0.50 – 0.69	Easy item
0.70 – 1.00	Very easy item

As the item difficulty index approaches 0, the question becomes more difficult; as it approaches 1, the question becomes easier (Atilgan, 2012). In this study, items 7, 10, 12, 14, and 21 were difficult items because they fell between 0 and 0.29 on the item difficulty index. Items 3, 4, 5, 8, 9, 11, 15, 16, 17, 18, and 19 were moderately difficult because they fell between 0.30 and 0.49. Items 16, 17, 18, and 19 are examples of medium-difficulty items, while items 1, 2, 6, 13, and 20, which fall between

0.50 and 0.69, are examples of easy items. The study does not include any items with an index between 0.70 and 1.

Furthermore, Table 8 in the Findings section provides an item evaluation regarding the test's index.

Data Analysis

First, teacher candidates were coded from 1 to 126 as T1 (Teacher Candidate 1), T2 (Teacher Candidate 2), and so on, in the research application. Then, the obtained data were analyzed using the TestAn software package with permission from the developer, Aydın (2013). The discriminative and difficulty indices of all test items were determined separately, and the analyses and evaluations of these items are presented in the findings section. Additionally, the Kuder-Richardson 20 (KR-20) formula was employed, yielding a reliability coefficient of 0.71. It was concluded that the

reliability coefficient falls within the limits accepted in the literature and that the test is sufficiently reliable.

Since the three-tier diagnostic test has three different stages, different categories emerge. Therefore, Table 5 presents the possible responses and categories of the three-tier diagnostic test. These categories are further divided into four: scientific/correct knowledge, lack of knowledge, misconception, and lucky guess/lack of confidence.

Table 5
Possible Responses and Categories of the Three-Tier Diagnostic Test

Tier 1 Responses	Tier 2 Responses	Tier 3 Responses	Categories
Correct	Correct	Certain	Possesses Scientific (Correct) Knowledge
Correct	Correct	Not sure	Guess/Lack of Confidence
Correct	Wrong	Certain	Conceptual Error
Wrong	Correct	Certain	Conceptual Error
Wrong	Wrong	Certain	Conceptual Error
Correct	Wrong	Not sure	Lack of Information
Wrong	Correct	Not sure	Lack of information
Incorrect	Wrong	Not sure	Lack of information

The Kuder-Richardson 20 (KR-20) formula was used to calculate the reliability of the three-tier diagnostic test for simple machines as a result of the applications carried out. In this formula, responses to a test item are scored as 1 if correct and 0 if incorrect and are calculated accordingly (Büyüköztürk et al., 2009). For an item in the three-tier diagnostic test to receive one point—that is, to be considered correct—the individual taking the

test must answer the first two tiers correctly and indicate certainty in the third tier. If an individual answers any of the first two tiers incorrectly or is unsure in the third tier, the answer is considered incorrect and coded as 0 points (Çetinkaya & Taş, 2016).

Table 6 below provides the scoring table for the possible responses to the three-tier diagnostic test for simple machines.

Table 6
Scoring Scale for the Three-Step Diagnostic Test on Simple Machines

Tier 1 Responses	Tier 2 Responses	Tier 3 Responses	Score
Correct	Correct	Confident	1
Correct	Correct	Not sure	0
Correct	Wrong	Not sure	0
Wrong	Correct	Certain	0
Wrong	Wrong	Confident	0
Correct	Wrong	Not sure	0
Wrong	Correct	Not sure	0
Wrong	Wrong	Not sure	0

According to Çepni (2007), "reliability" is a term used to describe how often findings can be repeated. In other words, it demonstrates the consistency of multiple measurement results for a specific quality of the measurement tool. The reliability coefficient takes a value between 0 and 1; as it approaches 1, reliability is high, and as it approaches 0, reliability is low. However, in the KR-20 formula, the test items may not have the same level of difficulty. When using the KR-20 formula for knowledge tests with a small number of items (10-15), a value as low as 0.50 can confirm the test's reliability (Şencan, 2005). The reliability value of the three-tier diagnostic test for simple machines in the study was determined using the KR-20 formula based solely on an analysis of the first-tier questions and was found to be 0.68. Additionally, when evaluated with the first two

tiers, the reliability was found to be 0.70. Finally, the reliability coefficient was found to be 0.71 when all three tiers were calculated together. In other words, the reliability of this test is above 0.70, indicating that it is reliable.

Findings

This section presents the item analysis results and interpretations for the three-tier diagnostic test items. The TestAn program output for the test items, i.e., the difficulty (p_j) and discrimination (r_{jx}) indices, are presented in Table 7 below. The table shows the evaluation of the "Tier 1 Score, Tier 2 Score Only, Score in Both Tiers (Tier 1 + Tier 2 Score), and Total Score (Score Calculated for All Three Tiers)" for the 21 questions in the three-tier diagnostic test (BMÜATT).

Table 7

Item Difficulty and Item Discrimination Indices of the Three-Tier Diagnostic Test Items for Simple Machines

Item No.	Tier 1 Score		Only Tier 2 Score		Score in Both Tiers		Total Score	
	p_j	r_{jx}	p_j	r_{jx}	p_j	r_{jx}	p_j	r_{jx}
1	.71	.53	.72	.56	.63	.62	.59	.47
2	.75	.50	.74	.53	.68	.65	.60	.67
3	.69	.21	.63	.50	.50	.71	.34	.50
4	.63	.38	.68	.47	.48	.62	.48	.50
5	.50	.65	.56	.65	.47	.71	.37	.62
6*	.72	.21	.66	.38	.72	.21	.59	.29
7	.59	.24	.59	.59	.31	.50	.25	.50
8	.34	.56	.66	.44	.44	.50	.40	.30
9	.69	.38	.59	.41	.38	.59	.34	.62
10*	.44	.24	.56	.35	.22	.27	.12	.12
11	.60	.68	.56	.65	.43	.68	.37	.62
12	.34	.38	.60	.56	.27	.41	.27	.35
13	.69	.38	.71	.56	.63	.32	.63	.33
14*	.41	.41	.35	.35	.29	.35	.28	.21
15	.47	.47	.52	.44	.53	.53	.41	.65
16	.48	.73	.48	.74	.46	.79	.32	.59
17	.84	.32	.53	.35	.41	.53	.40	.44
18	.63	.50	.62	.50	.57	.56	.47	.65
19	.68	.47	.59	.53	.49	.68	.37	.62
20	.65	.71	.59	.47	.65	.41	.50	.47
21*	.39	.08	.41	.18	.32	.41	.25	.27

*The four items indicated by * were not included in the final test because they had low discriminative power (index of 0.20 or less) when considering their total scores, or because the item with a borderline (between 0.20 and 0.29) needed to be revised.

The analysis revealed that items 1, 2, 6, 13, and 20 in the three-tier diagnostic test are easy ($PJ = 0.50-0.69$), while items 3, 4, 5, 8, 9, 11, 15, 16, 17, 18, and 19 are of medium difficulty ($PJ = 0.30-0.49$). Items 7, 10, 12, 14, and 21 are difficult ($PJ = 0.00-0.29$). Furthermore, the analysis revealed that test items 1, 2, 3, 4, 5, 7, 9, 11, 15, 16, 17, 18, 19, and 20 have very good discrimination power ($r_{jx} = 0.40$ or higher), while test items 8, 12, and 13 have fairly good discrimination power ($r_{jx} =$

0.30 to 0.39). Test items 6, 14, 21, and 10 have very low discrimination power ($r_{jx} = 0.19$ or lower). Based on these findings, items 1, 2, 3, 4, 5, 7, 9, 11, 12, 13, 15, 16, 17, 18, and 19 were included in the three-tier diagnostic test as they were discriminatory. Items 6, 10, 14, and 21 were not included as they were not discriminatory. The evaluation of the 21 questions in the three-tier diagnostic test is shown in Tables 8 and 9 below.

Table 8

Evaluation of the Items in the Three-Tier Diagnostic Test on Simple Machines in Terms of Difficulty Index

Item Difficulty Index (pj)	Evaluation Based on Item Number
0.00 – 0.29 (Difficult items)	7, 10, 12, 14, 21
0.30 – 0.49 (Moderately difficult items)	3, 4, 5, 8, 9, 11, 15, 16, 17, 18, 19
0.50 – 0.69 (Easy ones)	1, 2, 6, 13, 20
0.70 – 1.00 (Very easy)	-

Table 9

Evaluation of Simple Machines Three-Tier Diagnostic Test Items in Terms of Discrimination Index

Item Discrimination Index (r _{jx})	Evaluation by Item Number
0.40 and above	1, 2, 3, 4, 5, 7, 9, 11, 15, 16, 17, 18, 19, 20
Between 0.30 and 0.39	8, 12, 13
Between 0.20 and 0.29	6, 14, 21
0.19 and below	10

The item difficulty values for the total scores of the three-tier diagnostic test for simple machines range from 0.12 to 0.63. The average of these values is 0.40. In other words, these specified value ranges and the resulting average indicate that the test is of medium difficulty. The item discrimination values of the three-tier diagnostic test for simple machines range from 0.12 to 0.67. The average of these values is 0.47, which is considered very good.

Results, Discussion, and Recommendations

Within the scope of the research, a multiple-choice test (the Simple Machines Three-Tier Diagnostic Test) was developed to identify misconceptions that science teacher candidates hold about the topic of "Simple Machines" in Physics-1. Validity and reliability analyses were performed. Taking into account

the Science Teacher Education Undergraduate Program specified by the Council of Higher Education (YÖK) and expert opinions, the three-tier diagnostic test was developed, and the necessary adjustments were made based on the results of the pilot application. The test was administered to science teacher candidates, and the data obtained were analyzed. The analyses yielded the difficulty and discrimination indices of the test items, which were used to determine which items to include and exclude from the final test.

Three-tier diagnostic tests allow us to determine whether answers are correct or incorrect for the right or wrong reasons, respectively (Arslan et al., 2012; Bozdağ, 2017; Özden & Yenice, 2017). Progressing from single-tier tests to three-tier tests decreases the percentage of correct answers and

misconceptions. Furthermore, in three-tier tests, students are less likely to solve questions based on their estimated chance of success (Çiğdemoglu & Arslan, 2017; Elmas & Pamuk, 2021; Özden & Yenice, 2017). Examining numerous studies on this subject reveals that three-tier tests are more advantageous than single- or two-tier tests (Eryılmaz & Sürmeli, 2002; Halloun & Hestenes, 1995; Kirbulut & Geban, 2014; Mellyzar, 2021; Peşman & Eryılmaz, 2010; Taşlıdere, 2016). In other words, unlike single-tier diagnostic tests, three-tier diagnostic tests determine the level of concept learning and identify knowledge gaps, misconceptions, and lack of confidence.

The reliability coefficient reflects the internal consistency of a measurement tool (Tan, 2006). In this study, the reliability of the three-tier diagnostic test for simple machines, calculated using the KR-20 formula, was found to be 0.71. Different reliability values are accepted in the literature. For example, Kane (1986, p. 221) states that reliability coefficients above 0.50 are acceptable, while Tavşancıl (2006) states that the reliability coefficient should be 0.60 or above and Büyüköztürk (2013) states that it should be 0.70 or above. A review of the literature reveals that reliability coefficient values (KR-20 and KR-21) in science education test development studies are 0.60 or higher, which is consistent with the results of this study (Çakır & Aldemir, 2011; Çetinkaya, 2019; İspir, 2020; Orduhan & Çakır, 2023; Özcan, Çetinkaya, & Arık, 2021; Özden & Yenice, 2017; Özmen & Sever, 2024; Peker & Taş, 2019; Yavaş, 2019). In other words, the reliability value obtained in this study (0.71) is consistent with those reported in previous studies.

The item difficulty index in tests is expressed as the percentage of individuals in the tested population who answered the items correctly. As an item's difficulty index increases, the item is considered easier; as the index decreases, the item is considered more difficult. The item discrimination index is the ability of a test item and its distractors to distinguish between individuals who know and do not know the quality that the test aims to measure; in other words, it is the test's ability to discriminate. In prepared tests, higher discrimination indicates higher reliability (Baştürk, 2013). The item discrimination index

helps distinguish whether an individual who answers a question correctly possesses the required knowledge or skill. An item with an item discrimination index that can take values between -1 and +1 is suitable for use in the test when it approaches +1. The item difficulty index ranges from 0 to 1; the closer the value is to 0, the more difficult the item, and the closer the value is to 1, the easier the item. During the item analysis process, statistical and content-based criteria determine the level of item discrimination: Items with a discrimination index below 0.19 are excluded from the test or completely revised. Items between 0.20 and 0.29 are borderline and may be included in the test after revision. Items between 0.30 and 0.39 may be included in the test with minor revisions. Items with a discrimination index greater than 0.40 function very well and may be included in the test as is (Atılğan et al., 2015).

In this study, the item difficulty values of the three-tier diagnostic test for simple machines ranged from 0.12 to 0.63, with an average of 0.40. These value ranges and averages indicate that the test is of medium difficulty. According to Büyüköztürk et al. (2009), items with an item difficulty index between 0.40 and 0.59 are considered medium. According to Atılğan et al. (2015), items with an item difficulty index between 0.30 and 0.49 are also considered medium. A review of the literature on test development in science education found that the item difficulty index is 0.30 or higher (Avcı, Acar-Şeşen, & Kırbaşlar, 2018; Ayvacı & Durmuş, 2016; Çakır-Yıldırım & Karaarslan-Semiz, 2022; Çakır & Aldemir, 2011; Çetinkaya & Taş, 2016; Demir & Akarsu, 2014; Gülmez-Güngörmez & Akgün, 2018; Küçükkeskin & Kılıç, 2024; Özden & Yenice, 2017; Peker & Taş, 2019; & Saraç, 2018). This study found that the average item difficulty index of the test was 0.40, indicating moderate difficulty, consistent with previous studies.

The discriminant indices of the items in the three-tier diagnostic test for simple machines ranged from 0.12 to 0.67, with an average of 0.47 (considered very good). Items with a discrimination index below 0.19 cannot be included in the test and must be revised, while items with an index between 0.20 and 0.29 are borderline and may be included in the test after revision. However, to have high discriminability, a test must include items with

a discriminability index of 0.30 or above (Crocker & Algina, 2006). Thus, four items (6, 10, 14, and 21) with discrimination indices below 0.30 were excluded from the test. A review of the literature revealed that studies related to test development in science education had an average item discrimination index of 0.40 or higher (Çetinkaya & Taş, 2016; Çiğdemoglu & Arslan, 2017; Gülmez-Güngörmez & Akgün, 2018; Kargın & Gül, 2021; Özcan, Koca, & Söğüt, 2019; Özden & Yenice, 2017; Özmen & Sever, 2024; Peker & Taş, 2019; Saraç, 2018; Şimşek & Şahin-Çakır, 2025; Tüncaz & Şahin-Çakır, 2025). In this study, the average item discrimination index of the test was 0.47. In other words, the item discrimination was very good. This result is consistent with those of previous studies.

The three-tier diagnostic test initially consisted of 21 items. As a result of the applications and analyses performed, it was revised and reduced to 17 items, thus reaching its final form. The final version of the test is expected to be a useful resource for faculty members involved in training science teacher candidates and researchers working in this field.

Based on the results obtained from this study, the following recommendations can be made:

- In this research, the three-tier diagnostic test was used. Although three-tier diagnostic tests are better than two-tier diagnostic tests, they are determined by asking a single certainty question related to the main question and the justification section. It may be better if they are asked once more whether they are certain about the main question and the justification section (twice in total), in other words, if it is made a four-tier diagnostic test.
- In this study, the three-tier diagnostic test was administered to 126 science teacher candidates, and its reliability value was found to be 0.71. The reliability of the test can be tested by applying it to different study groups.
- In this study, four items were removed from the test. These removed items can be completely revised and reapplied to the same group to calculate reliability.

- In this study, a three-tier diagnostic test was developed for the topic of "Simple Machines." Three-tier diagnostic tests can also be developed for other scientific topics. When developing a three-tier diagnostic test, topics that students find difficult can be selected.
- In this study, the three-tier diagnostic test was prepared for science teacher candidates. Instead of science teacher candidates, a three-tier diagnostic test related to simple machines can be prepared for middle schools or high schools.

Ethics statement

This study was approved to comply with science ethics by Ondokuz Mayıs University's Ethics Committee of Social Sciences and Humanities (Research ethical authorisation number: 2024-798). Throughout the data collection, analysis, and reporting processes, utmost attention was paid to ethical principles regarding human privacy, data security, and confidentiality. Informed consent was obtained from all participants, and they were assured that their participation was voluntary.

Acknowledgments

The research presented herein emanates from the master's thesis authored by the primary investigator under the guidance and supervision of the second author.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflict of interest

None

References

- Acar, B., Korkmaz, Ö., Çakır, R., Erdoğan, F. U., & Çakır, E. (2019). Eğitsel robot setleri ile fen ve teknoloji dersi basit makinalar konusunun ortaokul 7. sınıf öğrencilerinin STEM beceri düzeylerine ve derse dönük tutumlarına etkisi [Educational robot sets with science and technology course basic

- machinery of the secondary school 7th class students' stem skill levels and the effect of the lesson attitudes]. *Eğitim Teknolojisi Kuram ve Uygulama*, 9(2), 372-391. <https://doi.org/10.17943/etku.518215>
- Aci-Özkan, D. (2024). *Türkiye’ de fen eğitiminde kavram karikatürlerinin kullanıldığı araştırmaların analizi [Analysis of studies using concept cartoons in science education in Turkey]* (Tez No. 850026) [Yüksek lisans tezi, Manisa Celal Bayar Üniversitesi]. YÖK Tez Merkezi.
- Amir, R., & Tamir, P. (1994). In-depth analysis of misconceptions as a basis for developing research-based remedial instruction: The case of photosynthesis. *The American Biology Teacher*, 56(2), 94-100. <https://doi.org/10.2307/4449760>
- Arslan, H. O., Çigdemoglu, C., & Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers’ misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International Journal of Science Education*, 34(11) 1667–1686. <https://doi.org/10.1080/09500693.2012.680618>
- Atılgan, H. (2012). *Ölçme ve değerlendirme [Measurement and Evaluation]*. Yediiklim Yayınları.
- Atılgan, H., Kan, A., & Doğan, N. (2015). Eğitimde ölçme ve değerlendirme. H. Atılgan (Ed.), *Test geliştirme* (s.316-348). Anı Yayıncılık.
- Avcı, D. E., Kara, İ., & Karaca, D. (2012). Fen bilgisi öğretmen adaylarının iş konusundaki kavram yanlışları [Misconceptions of science teacher candidates about work]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 31(1), 27-39. <https://dergipark.org.tr/pub/pauefd/issue/11112/132850>
- Avcı, F., Acar-Şeşen, B., & Kırbaslar, F.G. (2018). Maddenin yapısı ve özellikleri ünitesine yönelik iki aşamalı teşhis testinin geliştirilmesi [The Development of Two-Tier Concept Test for The Structure and Properties of Matter Unit]. *Kastamonu Education Journal*, 26(4), 1007-1019. <https://doi.org/10.24106/kefdergi.434239>
- Ayas, A., Çepni, S., Johnson, D., & Turgut, M.F. (1997). *Kimya öğretimi [Teaching of chemistry]*. YÖK / Dünya Bankası Eğitimi Geliştirme Projesi Hizmet Öncesi Öğretmen Eğitimi Yayınları.
- Aydın, A. (2013). *Çoktan seçmeli ölçme sonuçlarının bilgisayar yardımıyla analizi [Computer aided analysis of multiple choice test results]*. (Tez No. 336968) [Yüksek lisans tezi, Afyon Kocatepe Üniversitesi]. YÖK Tez Merkezi.
- Ayktulu, I., & Şen, A. İ. (2012). Üç aşamalı test, kavram haritası ve analogi kullanılarak lise öğrencilerinin elektrik akımı konusundaki kavram yanlışlarının belirlenmesi [Determination of secondary school students’ misconceptions about the electric current using a three tier test, concept maps and analogies]. *Eğitim ve Bilim Dergisi*, 37(166), 275-288.
- Ayvacı, H. Ş., & Devocioğlu, Y. (2009, Mayıs). *İlköğretim öğrencilerinin iş-güç-enerji konusunda sahip oldukları yanlış anlamalar [Misconceptions of primary school students about work, power, and energy]*, Sözlü bildiri sunumu]. First International Congress of Educational Research, Çanakkale.
- Ayvacı, H. Ş., & Durmuş, A. (2016). Bir başarı testi geliştirme çalışması: Isı ve sıcaklık başarı testi geçerlik ve güvenilirlik araştırması [An achievement test development study: heat and temperature achievement test validity and reliability research]. *Ondokuz Mayıs University Journal of Education Faculty*, 35(1), 87-103. Doi: 10.7822/omuefd.35.1.8
- Ayvacı, H. Ş., Özsevgenç, T., & Cerrah, L. (2004). Yıldırım kavramının farklı yaş grubundaki öğrencilerde gelişimi [The development of the concept of lightning in students of different age groups]. *Gazi Üniversitesi Kastamonu Eğitim Dergisi*, 12(2), 351-360.
- Baştürk, S. (2013). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Vize Yayıncılık.
- Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması [Measurement in education and psychology: Classical test theory and its applications]*. Pegem Akademi.
- Bilgin, A. K., & Kala, N. (2018). Fen bilimleri konularının günlük hayattaki yeri dersinin öğretmen adaylarının kavram ile bağlam ilişkisini oluşturabilmeleri üzerine etkisi [The effect of the lesson on the place of science topics in daily life on teacher candidates' ability to establish the relationship between concepts and context]. In *ERPA 2018: International Congresses on Education* (s. 70-76). Educational Researches and Publications Associations.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, the classification of educational goals. Handbook; cognitive domain*. (s. 141-225). New York, David McKay Company Inc.
- Bostan, A. (2008). *Farklı yaş grubu öğrencilerinin astronominin bazı temel kavramlarına ilişkin düşünceleri [Different age group students ideas about some basic astronomy*

- concepts] (Tez No. 237667) [Yüksek lisans tezi, Balıkesir Üniversitesi]. YÖK Tez Merkezi.
- Boz, Y. (2005). İlköğretim ikinci kademe ve ortaöğretim öğrencilerinin yoğunlaşma konusundaki kavram yanlışları [Second level primary education and secondary education students' misconceptions about the condensation concept]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 28(28), 48-54. <https://dergipark.org.tr/en/pub/hunefd/issue/7808/102419>
- Bozdağ, H. (2017). Üç aşamalı kavramsal ölçme aracı ile öğrencilerin sindirim sistemi konusundaki kavram yanlışlarının tespiti [Determining the misconceptions of students on digestive system by using 3-tier conceptual measuring scale]. *Bartın Üniversitesi Eğitim Fakültesi Dergisi*, 6(3), 878 – 901. <https://doi.org/10.14686/buefad.308999>
- Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı [Data analysis handbook for the social sciences]*. (8. Baskı). Pegem Akademi.
- Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2009). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Pegem Akademi.
- Cerit-Berber, N., & Sarı, M. (2009). Kavramsal değişim metinlerinin iş, güç, enerji konusunu anlamaya etkisi [The impact of conceptual change texts on understanding work, power, and energy]. *Selçuk Üniversitesi Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 27, 159-172.
- Colletta, A.T., & Chiappetta, E.L. (1989). *Science introduction in the middle and secondary schools* (2nd Ed.). Ohio, USA: Merrill Publishing Company.
- Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Fort Worth, TX, Harcourt College.
- Çakır, M., & Aldemir, B. (2011). İki aşamalı genetik kavramlar tanı testi geliştirme ve geçerlik çalışması [Developing and validating a two tier mendel genetics diagnostic test]. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 8(16), 335-353. <https://dergipark.org.tr/tr/pub/mkusbed/issue/19554/208352>
- Çakır-Yıldırım, B., & Karaarslan-Semiz, G. (2022). Üç aşamalı ekolojik ayak izi tanı testinin Türkçe'ye uyarlanması [Adaptation of a three-tier ecological footprint diagnostic test to turkish]. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, 23(2), 1588-1638. Doi: 10.29299/kefad.979056
- Çepni, S. (2007). *Araştırma ve proje çalışmalarına giriş [Introduction to research and project work]*. Celepler Matbaacılık.
- Çetinkaya, İ. (2019). *Basit makineler ünitesi ile ilgili geliştirilen düşünce deneyi etkinliklerinin 8. sınıf öğrencilerinin kavramsal anlamalarına etkisi [The effect of thought experiment activities on 8th grade students' understanding conceptual understanding of simple machines]* (Tez No. 585933) [Yüksek lisans tezi, Aksaray Üniversitesi]. YÖK Tez Merkezi.
- Çetinkaya, M., & Taş, E. (2016). “Vücudumuzda Sistemler” ünitesine yönelik üç aşamalı kavram tanı testi geliştirilmesi [Developing a three tier concepts diagnostic test on the outcomes of “The Systems in Our Body” unit]. *Ordu Üniversitesi Sosyal Bilimler Enstitüsü Sosyal Bilimler Araştırmaları Dergisi*, 6(15), 317-330. <https://dergipark.org.tr/tr/pub/odusobiad/issue/27575/290210>
- Çiğdemoglu, C., & Arslan, H. Ö. (2017). Atmosfer ile ilgili çevre problemleri konularında kavram yanlışlarını tespit eden üç aşamalı tanı testinin Türkçeye uyarlanması [Adaptation of a three-tier diagnostic test identified misconceptions on the atmosphere related environmental problems to Turkish]. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 14(1), 671-699. <http://dx.doi.org/10.23891/efdyu.2017.26>
- Dağdalan, G., & Taş, E. (2017). Simülasyon destekli fen öğretiminin öğrencilerin başarısına ve bilgisayar destekli fen öğretimine yönelik tutumlarına etkisi [Effect of simulation – based science teaching on students' achievement and attitudes towards computer – based science teaching]. *Fen Bilimleri Öğretimi Dergisi*, 5(2), 160-172. <https://dergipark.org.tr/tr/pub/fbod/issue/71996/1158045>
- Demir, N., & Akarsu, B. (2014). Modern fizik konuları ile ilgili kavram testi geliştirilmesi ve uygulanması: Modern fizik kavram testi (MKFT) [Development and implementation of a concept test related to modern physics topics: Modern physics concept test (MPCT)]. *Journal of European Education*, 4(2), 39-51. Doi:10.18656/jee.85816
- Diken, E. H. (2018). Fen bilgisi öğretmenleri ile 8. sınıf öğrencilerinin Temel eğitimden ortaöğretime geçiş (TEOG) sınavındaki kaygılarına yönelik görüşleri (Kars ili örneği) [The opinions of science teachers and 8th grade students regarding TEOG exam anxiety (the case of kars province)]. *İnsan ve Toplum Bilimleri Araştırmaları*

- Dergisi*, 7(2), 718-741.
<https://doi.org/10.15869/itobiad.376591>
- Ebel, R. L. (1995). *Measuring educational achievement*. New Jersey: Prentice-Hall Education Series.
- Elmas, R., & Pamuk, S. (2021). Öğretmen adaylarının kavram yanlışlarının üç aşamalı kavram yanlışlığı testi ile belirlenmesi [Determining misconceptions of prospective teachers with the three-tier misconception test]. *Afyon Kocatepe Üniversitesi Sosyal Bilimler Dergisi*, 23(4), 1386-1403.
<https://doi.org/10.32709/akusosbil.916063>
- Erkuş, A. (2017). *Davranış bilimleri için bilimsel araştırma süreci [The scientific research process for behavioral sciences]* (5. Baskı). Seçkin Yayıncılık
- Ersoy, E., & Bayraktar, G. (2018). İlkokul 4. sınıf matematik dersi “Ondalık Gösterim” alt öğrenme alanına ilişkin başarı testi geliştirilmesi [Development of achievement test related to sub-learning of decimal projection in math class of 4th grade in primary school]. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, (46), 240-266.
<https://dergipark.org.tr/tr/pub/deubefd/issue/41774/407618>
- Eryılmaz, A., & Sürmeli, E. (2002). Üç-aşamalı sorularla öğrencilerin ısı ve sıcaklık konularındaki kavram yanlışlarının ölçülmesi [Measuring students' misconceptions about heat and temperature using three-tier questions]. *Journal of Turkish Science Education*, 14(1), 110-126.
<https://users.metu.edu.tr/eryilmaz/TamUcBaglanti.pdf>
- Gülçiçek, Ç., & Yağbasan, R. (2004). Sarmal yay sisteminde mekanik enerjinin korunumu konusunda öğrencilerin kavram yanlışları [Students' misconceptions about the conservation of mechanical energy in a coil spring system]. *Milli Eğitim Dergisi*, 163, 144-154.
https://dhgm.meb.gov.tr/yayimlar/dergiler/milli_egitim_dergisi/163/gulcicek.htm
- Gülmez-Güngörmez, H., & Akgün, A. (2018). Ortaokul öğrencilerinin fen bilimleri dersindeki kuvvet ve enerji ünitesine yönelik akademik başarı testi geliştirme çalışması [The study of developing an academic success test aimed at the unit “force and energy” for secondary school learners in science courses]. *Diyalektolog-Uluslararası Sosyal Bilimler Dergisi*, 18, 85-99.
<http://dx.doi.org/10.22464/diyalektolog.218>
- Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, 53(11), 1043-1055. Doi:[10.1119/1.14030](https://doi.org/10.1119/1.14030)
- İspir, E. (2020). *Basit makineler ünitesinin öğretiminde kullanılan kavram karikatürlerinin 8.sınıf öğrencilerinin başarılarına ve kavramsal anlama düzeylerine etkisi [The effect of the concept cartoons used in the teaching of simple machines unit on the achievement and conceptual comprehension levels of 8th grade students]*. (Tez No. 642527) [Yüksek lisans tezi, Adıyaman Üniversitesi]. YÖK Tez Merkezi.
- İspir, E., & Aydın, M. (2020). Basit makineler ünitesinin öğretiminde kullanılan kavram karikatürlerinin 8. sınıf öğrencilerinin başarılarına ve kavramsal anlama düzeylerine etkisi [The effect of the concept cartoons used in the teaching of simple machines unit on the achievement and conceptual comprehension levels of 8th grade students]. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 1(38), 58-71.
<http://dx.doi.org/10.14582/DUZGEF.2021.154>
- Kane, M. T. (1986). The role of reliability in criterion-referenced tests. *Journal of Educational Measurement*, 23(3), 221-224. Doi:[10.1111/j.1745-3984.1986.tb00247.x](https://doi.org/10.1111/j.1745-3984.1986.tb00247.x)
- Kargın, P. D., & Gül, Ş. (2021). Altıncı sınıf vücudumuzdaki sistemler ve sağlığı ünitesine yönelik bir başarı testi geliştirilmesi [Development of an achievement test for sixth grade ‘body systems and health’ unit]. *Ihlara Eğitim Araştırmaları Dergisi*, 6(1), 1-26.
<https://doi.org/10.47479/ihead.729412>
- Kirbulut, Z. D., & Geban, O. (2014). Using three-tier diagnostic test to assess students' misconceptions of states of matter. *Eurasia Journal of Mathematics, Science & Technology Education*, 10(5), 509-521.
<https://doi.org/10.12973/eurasia.2014.1128a>
- Krathwohl, D. R. (2002). A Revision of Bloom’s taxonomy: An overview. *Theory Into Practice*, 41(4), 212-264.
https://doi.org/10.1207/S15430421TIP4104_2
- Küçükkeskin, E., & Kılıç, D. (2024). Hücre bölünmeleri konusunda öğrencilerin kavramsal anlamalarını belirlemeye yönelik iki aşamalı test geliştirilmesi [Development of a two-tier diagnostic test to determine students' conceptual understanding of cell divisions]. *Fen*

- Bilimleri Öğretimi Dergisi*, 12(1), 99-121.
<https://doi.org/10.56423/fbod.1380581>
- Laçin-Şimşek, C. (2019). *Fen öğretimde kavram yanlışları tespiti ve giderilmesi [Identifying and correcting misconceptions in science education]*. Pegem Akademi.
- Mellyzar, M. (2021). Analysis of understanding chemical bond concepts in students with three-tier multiple choice. *Journal of Educational Chemistry (JEC)*, 3(1), 53-66.
<https://doi.org/10.21580/jec.2021.3.1.7560>
- Orduhan, Y., & Çakır, Ç. Ş. (2023). Ortaokul 6. sınıf “ses ve özellikleri” ünitesine yönelik kavramsal anlama testi geliştirme çalışması [Middle school 6th grade conceptual understanding test development study for “sound and its characteristics” unit]. *Uluslararası Eğitim Bilim ve Teknoloji Dergisi*, 9(3), 138-178.
<https://doi.org/10.47714/uebt.1355916>