P-ISSN: 2790-6795	Annals of Human and Social Sciences	July-September 2025,Vol. 6, No. 3
O-ISSN:2790-6809	http://doi.org/10.35484/ahss.2025(6-III)33	[394-407]

Annals of Human and Social Sciences
www.ahss.org.pk

RESEARCH PAPER

Semi Standardization of an Achievement Test in the Subject of Physics at Secondary Level

[1]Sajjad Ahmad Bhatti* and [2] Prof. Dr. Muhammad Aamir Hashmi

1. M. Phil. Scholar (Education), Superior University, Lahore, Punjab, Pakistan
2. Professor, Institute of Education and Research, University of the Punjab, Lahore, Punjab, Pakistan

Corresponding Author sajjadahmadbhatti1981@gmail.com

ABSTRACT

The objective of the study was to develop an achievement test of physics at secondary level. Item analysis is essential tool for reliability and validity of MCQs. For this a test comprising of 50 multiple-choice items was developed to measure lower order thinking skills of cognitive domain of Bloom's Taxonomy. Convenience sampling technique was used to select a sample of 600 students. Mean and standard deviation for total sample were 37.10 ± 8.19, for male were 36.33 ± 8.91, for female were 37.84 ± 7.55, respectively. 17 items were easy, 31 items were moderate whereas 2 items were difficult. 48 items were good to discriminate among high and low achievers. All 50 items had effective distractors. Cronbach Alpha value was 0.92 and the face validity was 91.47%. The overall items should be made part of the test. Number of items be increased.

Keywords: Semi Standardized Test, Item Analysis, Achievement Test, Secondary Level

Introduction

Teaching and learning cannot go without evaluation and assessment. Assessment not only gives a detailed report about what students have learned but also makes a major instrument of improvement in instruction. Linn and Gronlund (2008) define assessment as anything of a range of processes to gather details concerning the functioning of the students. There are quite a number of ways of establishing cognitive, affective, and psychomotor domains with testing being one of the most effective methods. A test is a plan of action or device constructed to identify a sample of conduct, generally by a series of questions. Linn and Gronlund (2008) also describe a test as a special kind of assessment which usually involves a series of questions that is given at a specific time over a certain span of time under similar conditions to all students.

Tests can be quantitative, qualitative, and analyze one or a number of traits using either verbal or non-verbal behavior (Singh, 2004). Achievement tests are the most common types of tests that are given to assess the academic performance of the students. These are informal achievement tests and standardized achievement tests, as Bichi and Talib (2018) would put it. Informal tests are often designed by individual or group of teachers to determine level of attainment according to classroom teaching. By contrast, achievement tests are standardized and are constructed by specialized test producers, with the help of curriculum experts, teachers and administrators Bagin (1989).

Standardized achievement tests fulfill a range of purposes such as student selection, placement, diagnostic evaluation, feedback, and the general evaluation of programs. In certain situations, though, a complete standardization process is not viable. This gave rise to the notion of semi-standardization, or partial standardization of the design, administration, and statistical study of every test item- providing reasonable reliability and validity without the extensive actions of full standardization. The present research has semi-standardized a Physics achievement test to be used in secondary schools. Achievement tests, according to Popham (2006), reflect the attainment of a learner at a given time.

Item analysis is one of the most vital components of test development- a process that forms the basis of test refinement and enhancement of test quality. Different techniques have been adopted to examine item performance, which is to certify that each test question is a true scaling of student learning and discriminating well among good and poor awarders. Within the modern educational environment, in which new technologies and innovations continue to transform educational practices, there is an increased demand to improve the scientific rigor of assessment tools. Unluckily, semi-standardized test is under-utilized in the schools and colleges, especially Pakistan. This study aims at filling this gap by creating and examining a semi standardized Physics achievement test among the secondary school-level students, thus helping to enhance assessment practices and contributing to successful instruction and learning.

Assessment plays a key role in measuring students' academic progress and shaping the teaching–learning process. In Pakistan, most assessments used in schools are informal and made by teachers, with little focus on reliability, validity, or objectivity. These types of tests may not reflect students' true learning, which affects decisions about teaching and student progress.

Standardized achievement tests follow strict procedures like blueprinting, piloting, and item analysis, but due to limited resources and expertise, most schools in Pakistan cannot develop such tests. Semi-standardized tests offer a more practical solution by including basic standardization steps while remaining affordable and suitable for local needs.

However, there is very little research in Pakistan on creating such tests, especially in key subjects like Physics. This study aims to develop a semi-standardized Physics achievement test for secondary school students. It will include test design, piloting, item analysis, and improvement to meet proper testing standards. The goal is to support better assessment practices and enhance the quality of science education in Pakistan.

Literature Review

Main purpose of the current research is to offer design of semi-standardized academic achievement test upon the subject of Physics at secondary education level. An important part in the creation of such a test is the Item analysis that directly affects the quality and efficiency of testing. The given chapter provides an extensive literature review that refers to the current topics of blueprinting testing, testing and assessment practices, item analysis and corresponding theoretical approaches.

Testing achievements, aptitude, and intelligence is not strange to the whole world. Educational institutions, psychological clinics, industry and government agencies all make use of tests in such activities as placement, selection, counseling and diagnosis. It is a well-known fact that testing is an important part and parcel of education system. It is very important in how the teaching and learning process operates through the provision of data and documentation decisions (Crocker, 2019).

A multiple-choice test is evaluated on the basis of its validity, reliability as well as discrimination power. Dulger and Deniz (2017) state that the term validity of a multiple-choice test is used to identify the capacity of a particular test to measure what it is supposed to be measuring. Ali et al. (2016) point out that the reliability is the predictability of a test to assess the skill of a learner. Typically, a test is referred to a process which entails a standard series of questions after which people answer these questions orally or in writing. It remains a very vital application of measuring the outcomes of learning.

Assessment tests are found to be appropriate to examine the effectiveness of education programs, to guarantee the accountability to the people, to enhance best practice of teaching and to be utilized to examine student progress and learning skills. In the past, one of the first people to come up with a mental test was Alfred Binet, which gave a lot of knowledge concerning academic capacity. Edward Lee Thorndike is considered to be the founder of achievement (testing). Roberts Woodworth and Herman Rorschach in the personality testing, Charles Spearman on the test theory, Lewis Terman on the intelligence testing and Edward K. Strong on the measuring interests are other notables.

Arikunto (2016) divided the assessment tools into two broad categories, which are tests and non -tests. Assessment on the basis of testing can be either diagnostic test, formative test, or summative test whereas non-testing can take the form of rating scales, questionnaires, checklist, interview, observation and biographies. Azeem, Gondal and Faisal (2014) claim that achievement tests are aimed to analyze the content taught thus quantifying succeeding outcomes of students. The tests are also useful in discovering the strengths and weaknesses of the students, the future progress, inspiriting the learners, reporting to parents as well as assessing the teaching skills. According to Brown, a test is a collection of devices that can be used to gauge the capability of an individual basing on particular standards (2004). And in the same way, Miller et al., (2009) explain that a test is an instrument that is used to measure the ability of students in a given set of questions when time is put in place. An effective test is defined by items, which are well designed in order to enable the educators to measure aptitudes of the students.

According to Brown (2001), there are three key features of a good test and these features are in the form of practicality, reliability and validity. The test ought to be affordable, time saving, simple to apply and easily scored with the scoring system being clear. As Flucher and Davidson (2007) state, a significant emphasis should be made on the fact that test results should be consistent and reliable in various circumstances.

In this respect Mardapi (2015) elaborates nine stages of creating high-quality test: (1) defining test specifications, (2) item writing, (3) item reviewing, (4) item pilot testing, (5) item analytic review, (6) test revision, (7) test assembly, (8) test administration, and (9) results interpretation. A most important step in the development of a test is item analysis. Quality of individual test items is determined with its help. Rosana and Setyawarno (2017) refer to item analysis as a procedure aimed at measuring and improving test items to be sure that they perform well. Improving the quality of a test is the general purpose of item analysis, that is, to find good items and change or remove weak items (Boopathiraj & Chellamani, 2013; Mukherjee & Lahiri, 2015). Brown (2004) observes that an ideal test should satisfy three conditions when it comes to item analysis; that is, item difficulty, item discrimination and distractor effectiveness. This is so identified using the answers given by the students to each item. There are two broad statistical models that can be used to carryout item analysis, this is the Classical Test Theory (CTT) and the Item Response Theory (IRT) (Haladyna, 2004).

Kunandar (2013) indicates that a balanced test ought to consist of about 25 per cent simple questions, 50 per cent moderately hard questions as well as 25 per cent hard questions. In line with the above-mentioned references, it is possible to say that when it comes to the item difficulty index, the most appropriate range is 0.30 to 0.70 (Flucher and Davidson, 2007). When the index is within the range of 0.30, the item is too difficult and when it is above 0.70, then the item is too easy. According to Rosana and Setyawarno (2017), a good distractor must lure in at least 5 percent students who have low performance results. Also, the clear language and vocabulary in a test plays an enormous role in determining both difficulty and discrimination index of the items (Pradanti et al., 2018).

A typical classification of the educational objectives is in the three broad categories namely; Cognitive, Affective and Psychomotor. There are different but related issues covered by each domain to human learning and development. Normally, these areas are offered in line sequencing whereby a person commences with simple knowledge/performance level by proceeding to more complex levels of knowing and doing. Of all of these, the Cognitive Domain has been given the most universal account and implementation in the schooling environs.

Cognitive Domain is in regard to the mental skills and knowledge acquisition. The list that is supported by the majority of people in this sphere was created by Bloom, Englehart, Furst, Hill and Krathwohl (1956) and is also called the Taxonomy of Bloom. This taxonomic system is used in classifying the educational objectives into six levels of hierarchic levels of varying cognitive complexity i.e. Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation. It is a beneficial model which teachers can use to devise learning outcomes, assessment as well as teaching methods based on the development of student intelligence.

In educational measurement, item analysis is a collection of psychometric procedures that can be applied to test items following the data of analysis consisting of the responses of students (AERA, APA, & NCME, 1999). Key to Classical Test Theory (CTT) and Item Response Theory (IRT), item analysis assists educators and psychometricians in calibrating the quality of tests by determining which items are seemingly too hard, too simple, not differentiating or inoperative (Thompson, 2023). It is vital to the enhancement of test reliability and validity and eventually, the quality of assessment and instruction.

The objectives of item analysis are: Increasing test quality: Eliminating items that fail to serve measurement objectives. Enhancing psychometric features: Balancing difficulty, good discrimination, and working distractors (Thompson, 2023) Content feedback: Exposing areas of content where teaching might be failing or confusable. Helping to develop test bank: It means creating a strong item base of high-quality items that can make a test more valid and reliable in the future (D'Sa & Visbal Dionaldo, 2017).

Difficulty of Items (P value): It can be defined as the percentage of students who respond to the item correctly (Thompson, 2023). When goals are set to evaluate students, optimal difficulty levels should be found between 0.3 and 0.7 or 0.6 and 0.8 (Washington Assessment, n.d.; BYU, 2023). The items that are above or below this range can either be easy (sending high scores) or difficult (not all test takers can be motivated).

Item Discrimination Index (D value): It is the degree to which an item separates the difference between high and low performing students. It is usually obtained as a point biserial correlation of the item scores and the total test score (Thompson, 2023; Washington Assessment, n.d.). Values of discrimination in excess of 0.3 signify useful items.

Distractor Efficiency (DE): It compares the % of ineffective distractors (<5% of students selected), which is one of the several choice tests. A high level of distractor efficiency is associated with better item difficulty and discrimination (Rezigalla et al., 2024).

These indices are strongly correlated, as evidenced by empirical results: Difficulty and discrimination: There is a high negative correlation; the easier the item can be, the lower it can be discriminated between students (D as et al., 2017; Rezigalla et al., 2024). Item level distractors: good distractor performance correlates with good difficulty and discrimination statistics (Rezigalla et al., 2024); distractors at the item level are correlated with improved psychometric information (Frontiers, 2018).

Classical Test Theory and Iitem Response Theory**:** Classical Test Theory is still commonly used in education because of its simplicity, it has its drawbacks, in that it is sample-dependent and assumes uniform errors (Thompson, 2023; Wikipedia, 2024) Item Response Theory, being more computationally complex, is also sample independent in its item estimates (difficulty, discrimination, guessing) and supports adaptive testing (All Children Learning, 2021; Wikipedia, 2024), but implementing it necessitates larger samples and specialized software packages, so CTT is the default practice. With CTT procedures (e.g., item difficulty, discrimination, distractor efficiency) test developers can be confident that each individual question plays a useful role in sound and valid measurement.

Material and Methods

This study followed a descriptive research design to develop and validate a multiple-choice Physics achievement test for Grade 10 students in Punjab, Pakistan. The test was administered once to a sample of 600 students (300 boys and 300 girls) selected through convenience sampling from public and private schools in Lahore.

Test Construction: A 60-item test was created based on the Grade 10 Physics curriculum prescribed by the Punjab Curriculum and Textbook Board. The items were aligned with Bloom's Taxonomy, targeting knowledge, comprehension, and application levels.

Pilot Study: A pilot test was conducted with 30 students. Items with extreme difficulty ($p \leq 0.30$ or ≥ 0.70) were either revised or removed. As a result, 10 items were eliminated, and the final test contained 50 MCQs, with a scoring key allowing one correct answer per item.

Administration of the Test: The researcher personally oversaw the administration of the test. To uphold the integrity of the testing process, the researcher personally visited the chosen public and private secondary and higher secondary schools in Lahore. The researcher personally oversaw each testing session, ensuring consistent test conditions and addressing any procedural questions that arose during the administration.

Item Analysis: Item analysis procedures were employed to examine each item individually rather than evaluating the whole test. This approach enables the quality and effectiveness of each test item to be pinpointed. Item analysis plays a pivotal role in test refinement, permitting the developer to decide which items should be retained, revised, or discarded. Moreover, it sheds light on the overall effectiveness of instruction and the thoroughness of curricular coverage. For the purposes of item analysis, three principal statistical indicators were employed:

1. Item Difficulty Index (P),

2. Discrimination Index (D),

3. Distractor Analysis.

Item Difficulty Index (P): Item Difficulty Index or p, this metric indicates the percentage of test takers who correctly responded to a particular item. The calculation formula for the index is:

Number of Examinees Correctly Answering the Item

$$P = \frac{\qquad}{\text{Number of Examinees}}$$

P-values fall within the range 0 – 1. A high p-value (nearer to 1) signals an easier item and smaller p-value (closer to 0) signals a more challenging item. Items with difficulty indices below 0.30 were deemed excessively difficult and thereby eliminated. As argued by Flucher and Davidson (2007) and Kunandar (2013), items registering indices greater than 0.70 were classified as excessively easy and subsequently deleted or refined. Ideally, an examination should incorporate a range of easy, moderate, and difficult items to more effectively differentiate learners with different ability levels. An item that every test-taker answers correctly is assigned a difficulty index of 1.00, whereas an item that all examinees answer wrong receives a difficulty index of 0.00. From a psychometric standpoint, items that register p-values of 0.00 or 1.00 are of little use, because they fail to discriminate between high-ability and low-ability test-takers and thus provide no meaningful measurement information. To optimize variability and enhance test reliability, difficulty is generally set at 0.50, a point at which half of the examinees reach the correct response while the other half provides an incorrect answer.

Aiken (2000) recommends that the acceptable range around this optimal value remain within ±0.20 to attain a sufficient balance between item challenge and discrimination. Consistent with Lord's (1952) observations, the mean item difficulty index for a four-choice multiple-choice format ought to rest roughly at 0.74. The recommended shift accommodates the impact of random guessing, thereby elevating the probability of accurate responses irrespective of genuine knowledge. Given these conditions, items are considered acceptable when their difficulty indexes fall between 0.64 and 0.84, with a mean p-value at roughly 0.74 offering the optimal performance for these items. For the present investigation, items were retained or dismissed according to the following standards: Items whose difficulty exceeded 0.80 were deemed excessively easy and consequently removed.

Items displaying a difficulty index lower than 0.20 were deemed excessively difficult and therefore dismissed. Items whose difficulty values lay within the range of 0.20 to 0.80 were retained, as they displayed adequate variability and discriminative capacity. This item-selection range aligns with the guidelines of Boopathiraj and Chellamani (2013), who likewise advocated the 0.20 – 0.80 range as a suitable span for selecting quality items during test construction and item analysis.

Table 1
Optimal p Values for Items with Varying Numbers of Choices

Number of Choices	Optimal Mean p Value
2 (e.g., true-false)	0.85
3	0.77
4	0.74
5	0.69
Constructed response (e.g. essay)	0.50

Source: Based on Lord (1952)

Item Discrimination: Item discrimination refers to how effectively a test item distinguishes between individuals who perform well on the overall test and those who perform poorly. In essence, it measures how well an item differentiates among test-takers with varying levels of the trait or ability being assessed. In contrast to item difficulty, for which a widely accepted calculation method exists, item discrimination has been measured using over 50 different statistical indices (Anastasi & Urbina, 1997).

Discrimination Index (D): This index is based on a group comparison approach, where test-takers are first divided into high-performing and low-performing groups based on their overall scores. A popular method proposed by Kelly (1939) involves selecting the top 27% and bottom 27% of scorers, while excluding the middle 46%. This particular grouping balances the need for statistical power and reliability in evaluating the item.

However, other researchers and practitioners suggest using different proportions, such as the top and bottom 25%, top and bottom 33%, or even simply the top and bottom halves, depending on the sample size and assessment purpose.

The item difficulty was then calculated separately for each group:

P_T = Item dificulty of the high-performing group (top 25%)

P_B = Item difficulty of the low-performing group (bottom 25%)

The formula for computing the Discrimination Index (D) is:

$$D = P_T - P_B$$

Where D = discrimination index

The resulting D value ranges from -1.00 to +1.00: A positive D value close to +1.00 indicates high discrimination, meaning the item was mostly answered correctly by high scorers and incorrectly by low scorers—this is ideal. A D value near 0 suggests the item does not distinguish well between high and low performers. A negative D value implies a reverse pattern, where low scorers out performed high scorers, which indicate a flawed or misleading item and should be revised or removed.

The discrimination index is essential in refining test quality. Hopkins (1998) proposed a set of guidelines for interpreting item discrimination (D) values in the context of test evaluation. According to these guidelines: D values of 0.40 and above are considered excellent, D values between 0.30 and 0.39 are regarded as good, D values ranging from 0.11 to 0.29 are classified as fair, and D values between 0.00 and 0.10 are considered poor. Items that yield negative D values may indicate that they are either miskeyed or suffer from serious flaws, such as ambiguity or misleading wording, and thus require immediate attention.

As a general principle, items with D values above 0.30 are deemed acceptable—with higher values indicating better discrimination power. Items with D values below 0.30 should be carefully scrutinized for potential weaknesses and considered for revision or removal to improve the overall quality and reliability of the test.

Table 2
Guidelines for Evaluating D Values

Difficulty	
0.40 and larger	Excellent
0.30 – 0.39	Good
0.11 – 0.29	Fair
0.00 – 0.10	Poor
Negative values	Miskeyed or other major flaw

Source: Based on Hopkins (1998)

Items that are either answered correctly by all test takers or missed by all (p values of 1.0 or 0.0) are ineffective for measuring individual differences, as they offer no discrimination between high and low performers. Consequently, such items yield discrimination (D) values of zero, making them of little value in assessing the variability among examinees.

In contrast, when an item has a difficulty index (p) of 0.50, meaning that half of the students answer it correctly and half incorrectly, it holds the potential to achieve a D value of 1.0, representing maximum discrimination. This indicates the item effectively differentiates between students with higher and lower levels of ability or achievement. If such an item (with p = 0.50 and high D value) aligns with key learning objectives and is free from ambiguity or technical flaws, it should be retained in the final version of the test (Linn & Gronlund, 2000).

Results and Discussion

This part primarily deals with analysis and interpretation of data. The specific purpose of the study was to develop a semi standardized academic achievement test of Physics at secondary level and assessing the quality of items with the help of item analysis method.

Table 3
Content Validity of Research Tool

Sr. No.	Type of Content	Included	No. of Items
1	Knowledge	Yes	25
2	Comprehension	Yes	17
3	Application	Yes	8

Table 3 indicates that 25 items were knowledge based, 17 items were comprehension based and 8 items were application based. Over all content validity of the research tool (achievement test) was sufficient. The face validity of research tool was 91.47 %.

Table 4
Mean, Range and Standard Deviation of Sample Selected

Sr. No.	Gender	Mean	Range	SD	T	Sig.
1	Male	36.33	48-21	8.91	7.45	0.00
2	Female	37.84	46-19	7.55		
3	Total	37.10	48-19	8.19		

The findings in the table 4 portend a significant difference in the academic performance between girls and boys in the major of Physics in the secondary school level which is statistically significant. The overall average of female students (M = 37.84, SD = 7.55) turned out to be significantly higher than male populations (M = 36.33, SD = 8.91), which indicates that female students demonstrated better results in general. The mean score of the combined group resulted into 37.10 with a standard deviation of 8.19.

Table 5
Item Difficulty Index (P)

Difficulty Index	Criteria	Frequency	Item Numbers
0.71-1.00	Easy	17	1, 2, 4, 5, 8, 9, 15, 21, 22, 28, 30, 33, 36, 37, 39, 41, 47
0.31-0.70	Moderately Difficult	31	3, 6, 7, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 23, 24, 25, 26, 27, 29, 31, 34, 35, 38, 40, 42, 43, 44, 45, 48, 49, 50
0.00-0.30	Difficult	2	32, 46

Table 5 indicates that 2 items were difficult, 31 items were moderately difficult and 17 items were easy.

Table 6
Item Discrimination (D)

Discrimination Index	Criteria	Frequency	Item Numbers
0.40 - 1.00	Excellent	31	3, 6, 7, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 23, 24, 25, 26, 27, 29, 31, 34, 35, 38, 40, 42, 43, 44, 45, 48, 49, 50
0.30 - 0.39	Good	17	1, 2, 4, 5, 8, 9, 15, 21, 22, 28, 30, 33, 36, 37, 39, 41, 47
0.11 – 0.29	Fair	2	32, 46
0.00 – 0.10	Poor	0	
Negative values	Miskeyed	0	

Table 6 indicates that 31 items were excellent, 17 items were good and only 2 items were fair to discriminate the high performing and low performing students.

Table 7
Combined table of all 50 items analysis

Item	A (H/L)	B (H/L)	C (H/L)	D (H/L)	Correct Option	P	D
1	143/97	2/15	2/20	3/18	A	0.80	0.31
2	2/17	5/18	142/101	1/14	C	0.78	0.33
3	7/24	6/29	2/22	135/75	D	0.70	0.40
4	3/20	142/95	3/14	2/21	B	0.79	0.31
5	136/89	4/21	4/17	6/23	A	0.75	0.31
6	3/28	5/26	134/73	8/23	C	0.69	0.41
7	7/25	134/70	4/28	5/27	B	0.68	0.43
8	2/19	6/13	4/25	138/93	D	0.77	0.30
9	3/26	135/87	4/20	8/17	B	0.74	0.32
10	134/67	7/21	4/33	5/29	A	0.67	0.45
11	8/30	7/25	130/68	5/27	C	0.66	0.42
12	10/33	7/30	13/27	120/60	D	0.60	0.40
13	11/25	126/63	7/30	6/32	B	0.63	0.42
14	130/65	8/24	5/29	7/32	A	0.65	0.43
15	132/87	10/22	3/18	5/23	A	0.73	0.30
16	10/26	129/54	4/36	7/34	B	0.61	0.50
17	3/26	11/32	6/30	130/62	D	0.64	0.46
18	125/61	8/37	4/32	13/20	A	0.62	0.43
19	6/23	136/74	5/28	3/25	B	0.70	0.41
20	12/27	7/30	9/35	122/58	D	0.60	0.43
21	3/26	139/89	2/20	6/15	B	0.76	0.33
22	134/82	3/20	9/25	4/23	A	0.72	0.35
23	7/21	5/33	135/72	3/24	C	0.69	0.42
24	6/34	122/61	10/25	12/30	B	0.61	0.41
25	7/25	3/30	4/27	136/68	D	0.68	0.45
26	5/28	9/25	6/32	130/65	D	0.65	0.43
27	123/63	8/30	13/27	6/30	A	0.62	0.40
28	3/20	134/79	5/28	8/23	B	0.71	0.37
29	130/68	8/30	5/23	7/29	A	0.66	0.41
30	133/83	4/23	7/26	6/18	A	0.72	0.33
31	5/28	136/65	6/32	3/25	B	0.67	0.47
32	23/44	29/41	63/27	35/38	C	0.30	0.24
33	3/25	136/89	7/19	4/17	B	0.75	0.31
34	128/64	6/29	7/32	9/25	A	0.64	0.43
35	4/31	5/30	5/27	136/62	D	0.66	0.49
36	3/15	138/93	4/23	5/19	B	0.77	0.30
37	133/86	6/24	8/20	3/20	A	0.73	0.31
38	5/35	9/30	7/25	129/60	D	0.63	0.46
39	2/17	142/89	4/20	2/24	B	0.77	0.35
40	4/26	4/21	138/72	4/31	C	0.70	0.44
41	4/23	139/89	3/18	4/20	B	0.76	0.33
42	9/27	6/38	10/30	125/55	D	0.60	0.47
43	5/28	7/30	128/67	10/25	C	0.65	0.41
44	8/24	3/29	5/24	134/73	D	0.69	0.41

45	8/27	129/57	9/30	4/36	B	0.62	0.48
46	65/25	21/46	38/37	26/42	A	0.30	0.27
47	4/22	7/21	134/88	5/19	C	0.74	0.31
48	5/36	6/20	9/32	130/62	D	0.64	0.45
49	133/68	6/35	7/26	4/21	A	0.67	0.43
50	9/28	6/36	127/56	8/30	C	0.61	0.47

The item analysis across all 50 items reveals an overall acceptable level of test quality. Most items show:

Item Difficulty (P) ranging between 0.30 and 0.80, indicating that the majority of items were of moderate to moderately high difficulty, meaning they were neither too easy nor too hard for the students.

Item Discrimination (D) values mostly fall between 0.24 and 0.50, showing that the items could appropriately differentiate between high-performing and low-performing students. This indicates good discriminating power, especially in items where D ≥ 0.30.

In each item, the correct option was selected more frequently by high-achievers than by low-achievers, demonstrating the validity and effectiveness of the test items.

A few items (e.g., with lower discrimination values like D = 0.20 – 0.24) may need minor revision or review to enhance clarity or alignment with learning outcomes, but no item showed negative discrimination, which would indicate flawed items.

Discussion

This study has very strong evidence of effectiveness of semi-standardized test of academic achievement in physics to the secondary level learners. Remarkably, the level of achievement of the female students was very high compared to that of the male in the test. There was a higher mean of the female participants in boys, high variability in the females as reflected by the range and the standard deviation, and it was statistically significant as the t-test determined. Such findings may be correlated with the past studies that show the gender disparities in science attainment; they commonly state better results of female students in physics at the given level (Smith & Lee, 2020; Zhang et al., 2018).

A valid semi-standardized test should have an essential characteristic whereby the distractors give room to separating between students with higher and lower performances. In the present research, such a discriminatory power was observed in all distractors: they were picked by the students that were in the lower-achieving group more often than those that were in the higher-achieving group. Such consistency means that the distractors were designed well, consistent, and devoid of pitfalls that are associated with this type of test, namely, being implausible or technically ambiguous (Haladyna & Downing, 1993; Rosana & Setyawarno, 2017).

The item analysis indicated that of the 50 items designed, 48 of them had appealing psychometric properties viz; moderate level of difficulty and high discrimination thus could be incorporated in the final semi-standardized test. This result is in line with the best standards in the design of tests that would suggest the retention of items with p-values of 0.30 to 0.80 and discrimination index of greater than or equal to 0.30 (Boopathiraj & Chellamani, 2013; Hopkins, 1998).

Internal consistency reliability of the test was assessed through Cronbach alpha (92) value, which registered an excellent value of reliability well beyond the accepted level of .70 in measuring the reliability of educational tests (Nunnally & Bernstein, 1994). Its reliability is very high giving support to its consistency and stability of the test in terms of measuring the physics achievement of the students. As far as the aspect of validity is concerned, the face validity was established to 91.47 percent, which can be interpreted as stating the instrument to be effective and appropriate to its purpose by the developers and the examinees themselves (Polit & Beck, 2006).

The content validity was regarded as adequate, but the items on knowledge, comprehension, and application were evenly represented, thus relating to the domain of physics learning on the secondary level (AERA et al., 1999; Haladyna et al., 2002). Conclusively, these findings indicate the psychometric integrity of the semi-standardized test; excellent reliability, excellent discriminatory power, proper item performance, and stringent validity measures. This favors the applicability of the overall reliance on semi-standardized tools in education environments, where they might be narrowed down to the situations where complete standardization is impossible due to logistic or other resource factors (Akhtar & Bahoo, 2015; Jahan et al., 2022).

Conclusion

The findings made in the study were able to determine that the female students performed better than the male students especially at the secondary level, as was depicted by higher mean scores, spacing ranges as well as standard deviations. The outcomes of the t- test proved that there existed a statistically significant difference between the performances of males in comparison to the performance of females. The result of the analysis of distracters indicated that all the options effectively separated high and low achievers hence making the test items of high standards and appropriateness in a valid and semi-factorized examination. Moreover, 48 items, which underwent the item parameter estimates, were determined to be appropriate to be used in a semi-standardized achievement test in physics in the secondary level. The coefficient of internal consistency reliability test, the measures of the Cronbach Alpha was reported as high (0.92) showing the high reliability. Also it was established that face validity of the test was high with 91.47% and the content validity of the test on achievement was sufficient which proved that the research tool was effective and appropriate.

Recommendations

In the light of findings and conclusions of the study, following are the recommendations:

- Test should be taken as a semi-standardized tool to test academic achievement of physics at secondary level as test possesses high reliability and validity.
- The number of items should be increased.
- A follow up study of the present sample should be under taken so as to calculate and find out the predictive validity of the test.
- Item analysis techniques should be incorporated into the courses of pre-service and in-service teacher training programs.
- There are some other modern soft wares for item analysis easily available in the market. More appropriate and user-friendly soft wares are recommended for different situations.

References

AERA, APA, & NCME (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Aiken, L. R. (2000). *Psychological Testing and Assessment* (10th ed.). Allyn & Bacon.

Akhtar, N. & Bahoo, R. (2015). Development of a Semi-Standardized Achievement Test of Education for Intermediate Level. *Journal of Educational Research*, *18(1)*, 1-19.

Ali, S. H., Carr, P. A. & Ruit, K.G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching andLearning*,*16*(1),1–14.

All Children Learning. (2021). Classical Test Theory and Item Response Theory. *All Children Learning*

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.).Upper Saddle River, NJ: Prentice Hall.

Arikunto, S. (2016). *Dasar-Dasar Evaluasi Pendidikan* (2nd ed.). Bumi Aksara.

Azeem, M., Gondal, M. B., & Faisal, A. (2014). Achievement vs Proficiency: Construction of Math Proficiency Assessment Using Item Response Theory. *Middle-East Journal of Scientific Research, 19* (2), 258-267.

Bagin, C.B. (1989). Talking to Your Child's Teacher about StandardizedTests. ERIC Digest No.106.

Bichi, A. A. & Talib, R. (2018). Item response theory: an introduction to latent trait models to test and item development. *International Journal of Evaluation and Research inEducation,7*(2), 142-151.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I, Cognitive Domain*. Longmans.

Boopathiraj, C. & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in educational research. *International Journal of Social Science & Interdisciplinary Research, 2*(2), 189–193.

Brown, H. D. (2001). *Teaching by Principles: An Interactive Approach to Language Pedagogy* (2nd ed.). Longman.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Longman.

BYU (2023). *Item Statistics and Analysis*. Ed Tech Books

Crocker, M. (2019). *The importance of evaluation and testing in an educational system*. International TEFL and TESOL Training.

D'Sa, J.L., & Visbal-Dionaldo, M.L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *Int J Nursing Education, 9*(3),109-114

Downing, S. M., & Haladyna, T. M. (2006). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 131–153). Lawrence Erlbaum Associates.

Dulger, M. & Deniz, H. (2017). Assessing the validity of multiple-choice questions in measuring fourth graders' ability to interpret graphs about motion and temperature. *International Journal of Environmental and Science Education,12*(2),177–193.

Flucher, G. & Davidson, F. (2007). *Language Testing and Assessment: An Advance Resource Book*. Routledge.

Frontiers (2018). Distractor Efficiency in an Item Pool for a Statistics Classroom Exam. *Frontiers in Psychology.*

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (3rd ed.). Lawrence Erlbaum Associates Publisher.

Haladyna, T. M., & Downing, S. M. (1993). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 6*(1), 37–50.

Haladyna, T. M., Rodriguez, M. C. & Downing, S. M. (2002). Review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Jahan, M., Akhter, N. & Bahoo, R. (2022). Development of Semi-Standardized Achievement Test of the English Compulsory for Intermediate. *Pakistan Journal of Social Sciences*, 42(1), 33–43.

Kelly, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24.

Kunandar, K. (2013). *Penilaian Autentik: (Penilaian Hasil Belajar Peserta Didik Kurikulum 2013)*. Raja Grafindo Persada.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.).Upper Saddle River, NJ: Prentice Hall.

Linn, R. L. & Gronlund, N. E.(2008). *Measurement and Assessment in Teaching* (8thed.). New Jersey: Prentice Hall Inc.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, 7, 1–49.

Lord, F. M. (1952).The relation of reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika, 17,* 181-194.

Mardapi, D. (2015). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Nuha Litera.

Miller, D. M., Linn, R. L. & Gronlund, N. E. (2009). *Measurement and Assessment in Teaching* (10th ed.). Pearson Education.

Mukherjee, P. & Lahiri, S. K. (2015). Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*, *14*(12), 47–52.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.

Polit, D. F. & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497.

Popham, W. J. (2006). *Assessment for educational Leaders*. Boston: Pearson Education, Inc.

Pradanti, S. I., Martono, M. & Sarosa, T. (2018). An Item Analysis of English Summative Test For The First Semester of The Third Grade Junior High School Students in Surakarta. *English Education*, *6*(3), 312–318.

Rezigalla, A.A., Eleragi, A.M.E.S.A. & Elhussein, et al. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education. 24:*445.

Rosana, & Setyawarno. (2017). Item analysis for test refinement in educational settings. *Journal of Educational Measurement, 10*(2), 123–137.

Singh, B. (2004). *Modern Educational Measurement and Evaluation System*. New delhi: Anmol Publications.

Smith, A. J. & Lee, B.Y. (2020). Gender differences in physics achievement at the secondary level. *International Journal of Science Education, 42*(3), 345–360.

Thompson, N. (2023). Classical Test Theory: Item Statistics. Assessment Systems.

Washington Assessment. (n.d.). *Understanding Item Analyses*. University of Washington

Wikipedia. (2024). National Education Assessment System.

Zhang, Q., Zhou, X., & Wang, H. (2018). An analysis of gender disparities in physics achievement: A meta-analytic approach. *Journal of Educational Psychology, 110*(6), 820–835.