International Journal of Modern Education
EISSN: 2637-0905
IJMOE

Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

INTERNATIONAL JOURNAL OF MODERN EDUCATION (IJMOE)
www.ijmoe.com

GLOBAL ACADEMIC EXCELLENCE

THE DEVELOPMENT OF FORCE & MOTION ACHIEVEMENT TEST (FMAT) FOR FORM TWO STUDENTS

Putri Sathirah Saaban[1*], Nur Jahan Ahmad[2]

[1] School of Educational Studies, Universiti Sains Malaysia, Malaysia
Email: putrisathirah@student.usm.my
[2] School of Educational Studies, Universiti Sains Malaysia, Malaysia
Email: jahan@usm.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 24.10.2024
Revised date: 10.11.2024
Accepted date: 12.12.2024
Published date: 23.12.2024

To cite this document:

Saaban, P. S., & Ahmad, N. J. (2024). The Development Of Force & Motion Achievement Test (FMAT) For Form Two Students. *International Journal of Modern Education, 6* (23), 421-438.

DOI: 10.35631/IJMOE.623029

Abstract:

Force and Motion (FM) is one of the most challenging science concepts for students. However, the instruments for assessing students' understanding about this concept are limited. As such, this study focused on the development processes of an achievement test, specifically in this topic. Force and Motion Achievement Test (FMAT) was developed based on the Standard Curriculum of Secondary Schools (KSSM) for Science subject. It consists of 25 items, namely 22 objective items (Section A) and 3 subjective items (Section B). The FMAT was validated by 3 experts with more than 10 years experience. Consequently, 3 items were amended. A pilot study was conducted on 40 form two students, then followed by a reliability procedure through Internal Consistency with KR-20 for Section A and Split-Half method for Section B. The results from reliability analysis obtained an acceptable alpha coefficient value for both sections, which are 0.757 for Section A and 0.732 for Section B. In terms of difficulty index (p), there were 4 difficult items, 23 moderately difficult items, and 9 easy items. This findings indicate that the FMAT is valid and reliable to be utilised in any study involving the achievement test of Force and Motion topic. Future researchers are encouraged to perform item discrimination and distractor analysis to evaluate the items' ability in distinguishing among high and low achiever students, and how well the incorrect options divert students away from the correct answer. The implication of this study lies in providing a scientifically validated and reliable instrument for measuring students' understanding of Force and Motion topic, that help teachers identify learning gaps and enhance teaching strategies, ultimately improving the quality of science education.

Keywords:

Achievement Test, Difficulty Index, Force and Motion, Item Analysis, Reliability, Validity.

International Journal of Modern Education IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

Introduction

The national education system aims to elevate Malaysia's standing into the top third of international assessments, specifically in TIMSS and PISA. However, the average Science score of Malaysian students in TIMSS 2019 was 469, reflecting a slight decrease of 2 points from TIMSS 2015 (471) (MOE, 2020[b]). Besides, Malaysia's performance in PISA 2022 also indicates a decline, with the average scores of Malaysian students falling below the OECD average across all three literacy domains (MOE, 2023).Therefore, to achieve the goals set forth in the Malaysian Education Blueprint 2013-2025, it is essential to strengthen knowledge and assessment skills among teachers, in order to evaluate students' abilities to use and apply content knowledge. Hence, a variety of assessment strategies including both formative and summative approaches are employed to effectively measure and appraise students' academic progress and achievements. These methods not only provide insights into students' current understanding but also guide instructional practices and interventions to enhance educational outcomes (Haw et al., 2022).

To date, the educational system in Malaysia has undergone several transformations which aims to holistically equip students to compete in the global economy and society of the 21st century. Despite numerous transformations in Malaysia's educational system, there remains a significant challenge in developing effective and relevant methods to assess whether students are meeting the required standards (Narinasamy & Nordin, 2018; Chin et al., 2019; Hock et al., 2022; M. Yusop et al., 2024). Furthermore, developing a high-quality achievement test is a complex and challenging process. Therefore, it is essential for teachers to possess the necessary skills to effectively develop and analyse an achievement test (Haw & Sharif, 2020).

Achievement test is an evaluative tool designed to gauge the extent to which students have acquired new knowledge or skills (Nugba & Quansah, 2021). These tests are predicated on the educational principle that the learning process inherently leads to observable changes in an individual's behaviour. Consequently, the use of achievement tests in the educational setting can effectively pinpoint areas of difficulty that students may encounter during their learning journey and assess their preparedness to engage with new material (Esomonu & Eleje, 2020). Within the realm of educational research, instrument such as the Science Achievement Test are pivotal in gauging students' comprehension and mastery of scientific concepts.

The assessment of students' understanding in science, particularly in the topic of Force and Motion is essential for evaluating their conceptual grasp and problem-solving abilities. Studies have reported that students often harbor misconceptions about force and motion (Volfson et al., 2020; Bouzid et al., 2022; Bahtaji, 2023; Khoirunnisa et al., 2024) such as believing that a constant force is required to maintain motion or misunderstanding the relationship between force and acceleration (Khoirunnisa et al., 2024). For instance, many students think that if an object is moving, there must be a force in the direction of its motion (Bahtaji, 2023). Given these prevalent misconceptions, it's imperative to develop assessment tools that not only evaluate students' knowledge but also identify and address these misunderstandings.

The existing instruments may not adequately reveal students' misconceptions. This limitation results in significant gaps in diagnosing learning challenges and tailoring instructional strategies, thereby hindering both the teaching and learning processes. Addressing these shortcomings requires a critical review and refinement of existing assessment tools to ensure they can accurately capture the complexities of student understanding. In the context of this

International Journal of Modern Education IJMOE EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029
study, the achievement test for the topic Force and Motion is developed based on the components of physical science discipline (MOE, 2016). However, there aren't many studies in Malaysia involving the development of achievement test related to this concept that can be used to measure the understanding of form two students. This is because, most of the instruments developed are only suitable for measuring the understanding of students in the upper secondary level, as well as Science stream matriculation level (Ismail & Ayop, 2016; Shahari & Phang, 2023). Therefore, this study aims to address the following objectives:

1. To develop the Force and Motion achievement test (FMAT) that has satisfactory validity.
2. To evaluate the reliability of the Force and Motion achievement test (FMAT).
3. To analyse the psychometric properties of the FMAT items through item analysis.

Literature Review

Science Achievement Test

Achievement is frequently regarded as a crucial factor in education, serving as a benchmark for evaluating the effectiveness of teaching and learning processes. However, achievement test can either accurately or inaccurately indicate how well the students' learning. This is because the student's ability and the characteristics of items are the factors that affect a student's achievement in the examination (Suppiah Shanmugam et al., 2020; Haw et al., 2022). Thus, the development of high-quality achievement test requires substantial knowledge either in content area or in item development, to ensure the test items are reliable and valid.

According to MOE (2020[a]), the following principles, namely administration, ease of interpretation, comprehensiveness, validity, reliability and objectivity should be considered in the process of developing assessment instruments. As a result, the development of test items that are carried out in a planned and systematic manner can ensure that the test items are valid and reliable (Bekoe, 2023; Obilor & Omeke, 2024). Thus, it can be used as a measuring tool to assess how well students have achieved the learning outcomes.

In the Malaysian educational system, Science is one of the most significant disciplines for students to grasp. Nevertheless, low science achievement is still a problem in Malaysia (Halim & Meerah, 2016; MOE, 2020[b]; Phang et al., 2021; Badruldin & Alias, 2022). Moreover, previous studies have shown that among the most challenging topics in Science is the topic of Force and Motion (Ibrahim et al., 2019; Kamarrudin et al., 2020; Badruldin & Alias, 2022), which covers a wide range of concepts and applications in life. Based on the Standard Curriculum of Secondary Schools (*Kurikulum Standard Sekolah Menengah*, KSSM) for Science Form Two, the topic Force and Motion includes two Content Standards (SK) that must be mastered by students, namely Force and Effects of Force (MOE, 2016).

Although the concepts of Force and Motion have been introduced to students in Malaysia since primary school, most studies show that students' understanding of these concepts are still at a low level (Kamarrudin et al., 2020; Badruldin & Alias, 2022). It is undeniable that the understanding of Force and Motion concept is very important for students to grasp. This concept is highlighted across multiple levels of education in Malaysia, including both lower and upper secondary levels, particularly through subjects like Science (MOE, 2018[a]), and Physics (MOE, 2018[b]), as a continuation of the Science learning in form two. Therefore, to measure the student's understanding of the Force and Motion concept, a quality instrument

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

needs to be developed. However, most recent studies, especially in Malaysia, have focused on the development of instruments aimed at determining students' misconceptions about the concept, which adapting and modifying the instruments from the Force Concept Inventory (FCI) (Ismail & Ayop, 2016; Basran & Lajium, 2020; Murshed et al., 2020; Shahari & Phang, 2023). This clearly indicates that the instruments developed by previous researchers are not in line with the objectives of this study, that focus on the development of instruments that align with the content coverage in the KSSM for Science Form Two.

Classical Test Theory

Studies have also shown that many teachers in Malaysia used Classical Test Theory (CTT) in analysing test items (Abd Latif et al., 2016; Suppiah Shanmugam et al., 2020). CTT is widely used because it is a more practical method for developing quality test items (Suppiah Shanmugam et al., 2020) and apply simple theoretical assumptions that are easily fulfilled by test data (Hambleton & Jones, 1993; Champlain, 2010). The fundamental assumptions of CTT include: (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores from parallel tests are uncorrelated (Magno, 2009). The CTT assumes that each individual has a true score that would be obtained if there were no measurement errors. However, due to the imperfections in measurement, the observed score for an individual may differ from their true score, with this difference caused by measurement error (Kaplan & Saccuzzo, 2017).

The CTT has been applied to assess the reliability and various characteristics of instruments (Bichi & Embong, 2018; Haw & Sharif, 2020; Suppiah Shanmugam et al., 2020). It centres around three forms of scores: (1) the observed score, (2) the true score, and (3) the error score (Hambleton & Jones, 1993). This relationship can be described in the theoretical framework known as the 'Classical Test Model' which can be represented as:

$$X = T + E$$

where,
X = observed score (the actual score a student gets)
T = true score (the student's real ability or potential)
E = error score (random factors that affect the observed score)

By utilizing CTT, teachers can gain a deeper understanding of the quality of test items and effectively identify flawed items from a psychometric standpoint, particularly in terms of how students perceive the difficulty of those items (Suppiah Shanmugam et al., 2020).

Item Analysis

The common measure of item analysis statistics is item difficulty index. The advantages of the item analysis are not only to assess the quality of test items, but also to reflect the effectiveness of teaching and learning strategies (Mohd Noor, 2021).

Item Difficulty Index

The difficulty index (p) represents the frequency of students who correctly answered the item. A high difficulty index indicates that if a large number of students answered the item correctly, it can be classified as an easy item (Ebel & Frisbie, 1991). Conversely, a low index value means fewer students got the item right, making it a more difficult item.

International Journal of Modern Education
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

Methodology

Research Design

This study employed a quantitative approach, focusing on the systematic development of an achievement test instrument. This research design involves developing test items to assess students' understanding of the Force and Motion topic and followed by confirming their reliability and validity through statistical analysis and expert evaluation.

The development process of the Force and Motion Achievement Test (FMAT) instrument is divided into two phases, adapted from approaches suggested by Ong and Mohamad (2014), as well as Che Isa and Azid (2021). The first phase is the development of the items which involves five main steps namely (1) identifying the objectives of the instrument; (2) determining the content of the instrument; (3) prepare a Test Specification Table; (4) writing items; and finally (5) analysing items by 'subject matter experts' to determine the face validity and the content validity of the FMAT instrument. The second phase involves the process of verifying the psychometric aspects of the instrument by (1) conducting a pilot test; (2) determining the reliability value of the instrument; and (3) determine the item's difficulty index. The entire process of developing FMAT instrument is illustrated in Figure 1.

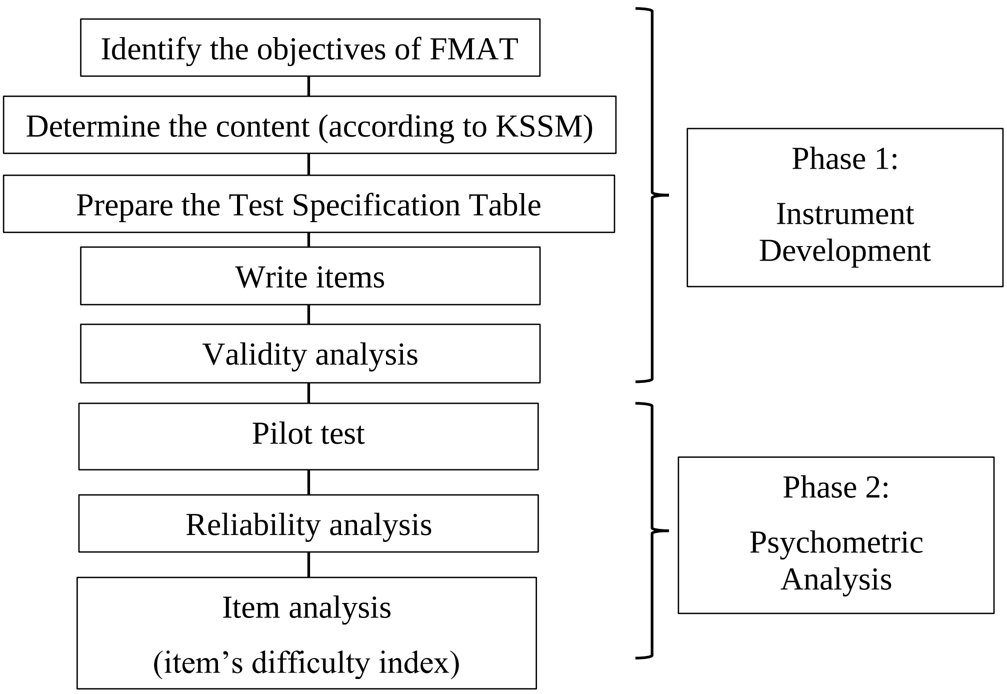


Figure 1: The Development Process of FMAT Instrument

To ensure that the FMAT instrument has high validity in terms of content validity and face validity, the FMAT instrument was given to experts who have skills in related fields. The validation process was carried out by three experts with more than 10 years of teaching experience, namely two Science teachers, and one SISC+ Science officer. In the context of this study, the experts needed to review the appropriateness of the language, the accuracy of the construct and the content, as well as the recommendations for necessary improvements. Meanwhile, for the face validity of this instrument, it was carried out by obtaining feedback from the experts, in terms of the suitability of the diagram, layout, language and the accuracy

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

of the terms used. Then, a pilot test was conducted to identify any problems or issues that may arise (Creswell & Creswell, 2018; Chua, 2012) in relation to the FMAT instrument, and determining whether there are items that need to be modified, discarded, or maintained in the FMAT instrument. Subsequently, from the data gathered through the pilot test, the researchers performed an item analysis using Classical Test Theory (CTT) to assess the difficulty index (p) of the items. This approach was applied to ensure that the selected items met the appropriate standards regarding difficulty and discrimination levels (Haw et al., 2022)

Sample

This study involved a total of 40 Form Two students in one of the schools in the district of Larut Matang & Selama, Perak. The selection of Form Two students was based on the suitability of the FMAT instruments that was developed using the topic of Force and Motion for Science Form Two. The instrument was administrated to all students in the pilot test to obtain the reliability value of the FMAT. The numbers of students involved (40 students) as a sample in this study is suitable for the purpose of the pilot test, which is between 25 and 100 people (Cooper & Schindler, 2011).

Instrument

The Expert Assessment Form (EAF) which has a scale of 1 to 4 was used to determine the validity for Force and Motion Achievement Test (FMAT). FMAT test contained 25 items. The EAF instrument uses 4-point Likert scale to avoid the existence of a neutral value or a midpoint value that makes it difficult to obtain a firm validity (Polit & Beck, 2006). The development of the FMAT instrument is guided by the KSSM for Science Form 2 (MOE, 2016) and covers the six levels found in the Revised Bloom Taxonomy (Anderson & Krathwohl, 2001). In addition, all items built also meet the predetermined difficulty level ratio, which is 5:3:2 which represents the Low: Medium: High level (MOE, 2022).

To measure students' mastery in the concept of Force and Motion, FMAT was built by adapting the format of the End of Academic Session Test (*Ujian Akhir Sesi Akademik*, UASA) for science subjects at the secondary school level (MOE, 2022). In the context of this study, the FMAT instrument is divided into two parts, namely Section A and Section B. Section A contains 20 multiple-choice items (MCQ), and 2 limited-response objective items. Meanwhile, Section B contains 3 subjective items. Subjective items are well-suited for assessing higher-order cognitive skills, such as analyzing, evaluating and creating, as they allow students to generate their own answers rather than choosing from predetermined options. It also enables students to provide detailed and insightful answers (Connor Desai & Reimers, 2019).

Data Analysis

The data analysis in this study was carried out using several statistical methods to determine the validity and reliability of the instruments. The validity of the FMAT instrument is determined based on the Content Validation Index (CVI). The CVI provides an average rating score for all items assessed by experts. Prior to CVI calculation, the relevance rating must be re-coded as '1' for a relevance scale of 3 or 4, and '0' for a relevance scale of 1 or 2 (Yusoff, 2019). The formula for calculating the i-CVI value is analysed according to the formula from Polit and Beck (2006), as follows:

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

$$\text{Item Content Validity Index (i-CVI)} = \frac{\text{The sum of each expert's score}}{\text{Total actual score}}$$

On the Expert Assessment Form, the experts are asked to give a score rating of 1 to 4 for each test item. Scores of 1 (Very Not Relevant) to 4 (Very Relevant) are the level of expert approval for an item that has been built. A comment and suggestion section are also provided for the purpose of improving items.

The reliability of the FMAT instrument was determined by using two main methods, namely the Kuder-Richardson-20 (KR-20) and Split-half. The KR-20 is used to measure the internal consistency of objective items, while the Split-half is implemented to measure the reliability of internal consistency of subjective items. The split-half reliability compares the performance between two halves of the test, typically divided by odd-even item splits.

Reliability is expressed numerically, usually as a coefficient ranging from 0.0 to 1.0 (Mills & Gay, 2016). A high coefficient (coefficient close to 1.00) signifies high reliability. Fraenkel and Wallen (1996) suggested that the reliability of items is acceptable if the alpha is within .70 and .99. In the context of this study, the collected data obtained from the pilot test was analysed using SPSS software version 26.0. In addition, the psychometric properties of each item in FMAT instrument were determined through item analysis, namely difficulty index.

The difficulty index (p) represents the frequency of students who correctly answered the item. It can be ranged from 0.0 to 1.0, and the formula to calculate the difficulty index for each item is given as follows (Mahjabeen et al., 2017):

$$\text{Difficulty index, } p = \frac{(H + L)}{N}$$

where,
H = number of students who answered correctly in the upper group
L = number of students who answered correctly in the lower group
N = total number of students in the upper and lower groups

In addition, the formula for calculating the difficulty index (p) of subjective items is analysed according to the formula from Suppiah Shanmugam et al., (2020), as follows:

$$\text{Difficulty index, } p = \frac{\sim (f)(x) - n\,(X_{min})}{n\,(X_{max} - X_{min})}$$

where,
f = number of students who obtained a specific score
x = scores of the item
~ f x = sum of ‘the number of students and their corresponding scores’
n = total number of students
X_{max} = maximum scores available for the item
X_{min} = minimum scores available for the item

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

A general guideline for the interpretation of item difficulty index values is shown in Table 1.

Table 1: Interpretation of Item Difficulty Index

Difficulty Index (*p*)	Interpretation
≤ 0.30	Difficult
0.31 ≤ 0.70	Moderately difficult
> 0.70	Easy

Source: Henning, 1987

To calculate the item difficulty index, the total test scores were first arranged in descending order from the highest to the lowest. The students were then split into groups, which the upper group representing the top 50% of scores, and the lower group representing the bottom 50%. However, the proportion of students in both groups can be either 27%, 33%, or 50%, according to the size of the students (Mohd Noor, 2021). In the context of this study, the pilot test was administered to only one class of students who consist of 40 students; hence the number of students was set at 20 for both the upper group and lower group. Following this, the number of correct answers in both the upper group (H) and the lower group (L) was counted for each item. Finally, the difficulty index was calculated for each item based on the students' responses.

Result

Description of FMAT Instruments

The development of FMAT instrument involved several important steps, including the preparation of the Test Specification Table (Reynolds et al., 2021). According to Irwing and Hughes (2018), Test Specification Table outlines the content areas and cognitive levels to be evaluated, ensuring that the test aligns with learning objectives and comprehensively covers the intended material. This structured approach enhances the validity and reliability of the assessment by providing a balanced representation of topics and skills, as well as to ensure that the number of questions covers the entire topic to be tested. Table 2 shows the description of the FMAT instruments which includes the number of items, item numbers based on topics and weightage for each sub-topics in more detail.

Table 2: Item Descriptions based on Sub-topics and Bloom Taxonomy Levels

Context	Revised Bloom Taxonomy						Number Of Item	Weightage (%)
	Remembering	Understanding	Applying	Analysing	Evaluating	Creating		
	L	L	M	M	H	H		
8.1: Force								
8.1.1 Elaborate and communicate about force.	1,3	2	-	-	25(b-i), 25(b-ii)	-	5	14
8.1.2 Explain that force has magnitude, direction and point of application.	-	4	5	-	-	-	2	6

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

8.1.3 Measure force in S.I. unit	-	25(a)	6	-	-	-	2	6
8.1.4 Explain with examples that every action forcé has an equal (same magnitude) reaction force but in the opposite direction.	21(a)	7,8	-	-	-	-	3	8
8.2 Effects of force								
8.2.1 Elaborate and communicate about the effects of force	-	9, 23(a)	21(b)	-	-	-	3	8
8.2.2 Elaborate and communicate about the effects of forcé.	-	10	11	-	-	-	2	6
8.2.3 Classify and solve problems on levers based on the position of fulcrum, load and effort.	-	12, 24(a)	13, 23(b)	24(b)	–	-	5	14
8.2.4 Explain and communicate about the moment of force.	-	-	14,15	-	23(c)	-	3	8
8.2.5 Carry out an experiment and communicate about pressure and its application in daily life.	-	16, 22(a)	-	-	24(c), 24(d).	-	4	11
8.2.6 Elaborate and communicate about gas pressure based on the kinetic theory of gas.	17	-	22(b)	-	-	-	2	6
8.2.7 Explain and communicate about the existence of atmospheric pressure and the effects of altitude on the magnitude of pressure.	-	18	19	-	25(c)	25(d)	4	11
8.2.8 Explain the effects of depth on liquid pressure.	-	20	-	-	-	-	1	2
Number Of Item	**18**		**11**		**7**		**36**	**-**
Weightage (%)	**50**		**30**		**20**		**-**	

Based on the UASA format, all items built must meet the prescribed difficulty level ratio of 5:3:2 (MOE, 2022) with reference to the 6 levels of cognitive domains in the Revised Bloom Taxonomy (Anderson & Krathwohl, 2001). The development of the item starts at the low level, namely the remembering level (C1) and the understanding level (C2); at the medium level, namely the applying level (C3) and the analysing level (C4); and at the high level, the evaluating level (C5) and the creating level (C6). In total, all 25 items that have been developed led to 100% weightage.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

Validity of FMAT Instrument

The FMAT instrument was reviewed by three experienced experts through Content Validation Index (CVI) to assess its accuracy, ensuring high content validity. Table 3 shows the results for the i-CVI value and the S-CVI value for the FMAT instrument.

Table 3: I-CVI & S-CVI for the Content Validity of FMAT Instrument

Item		Expert 1	Expert 2	Expert 3	Number of Approvals	i-CVI	
Section A							
1		1	1	1	3	1	
2		0	1	1	2	0.67	*
3		1	1	1	3	1	
4		1	1	1	3	1	
5		1	1	1	3	1	
6		1	1	1	3	1	
7		1	1	1	3	1	
8		1	0	1	2	0.67	*
9		1	1	1	3	1	
10		1	1	1	3	1	
11		1	1	1	3	1	
12		1	1	1	3	1	
13		1	1	1	3	1	
14		1	1	1	3	1	
15		1	1	1	3	1	
16		1	1	1	3	1	
17		1	1	1	3	1	
18		1	1	1	3	1	
19		1	1	1	3	1	
20		1	1	1	3	1	
21	(a)	1	1	1	3	1	
	(b)	1	1	1	3	1	
22	(a)	1	1	1	3	1	
	(b)	1	1	1	3	1	
Section B							
23	(a)	1	1	1	3	1	
	(b)	1	1	1	3	1	
	(c)	1	1	1	3	1	
24	(a)	0	1	1	2	0.67	*
	(b)	1	1	1	3	1	
	(c)	1	1	1	3	1	
	(d)	1	1	1	3	1	
25	(a)	1	1	1	3	1	
	(b)	1	1	1	3	1	
	(c)	1	1	1	3	1	
	(d)	1	1	1	3	1	
S-CVI/Ave = 0.97							

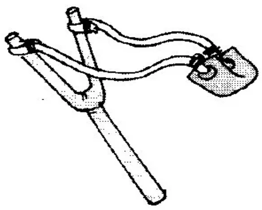
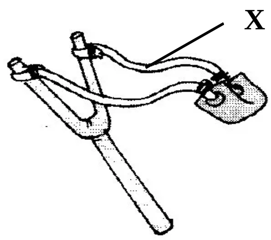
International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

Based on Table 3, the 'S-CVI / average' value obtained for the FMAT instrument is 0.97 which is higher than the S-CVI value of 0.80 proposed by Davis (1992). However, there are some FMAT items that have an i-CVI value of less than 1.00, as suggested by Polit and Beck (2006). Therefore, improvements were made to the three items that did not reach the proposed i-CVI value.

Items' Improvement Based on The Expert Recommendations

Based on the comments from experts, improvement was made to the items in terms of sentence structure and visual display. This modification is in line with the three components of the item's development, as proposed by Miller et al. (2013) namely stimulus, body of question (stem), visual and choice of answers. Overall, a total of 3 items have been improved based on the comments from expert, and no items have been removed. Table 4 shows the three original items that have gone through the process of modification and improvement based on experts' comments.

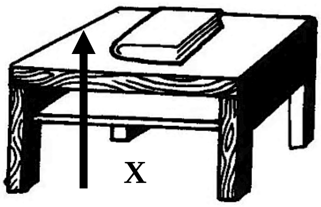
Table 4: Summary of the Items' Improvement for FMAT Instrument

No item	Original Item	Improvement Item
2	Berdasarkan rajah berikut, namakan jenis daya yang terlibat. *Based on the following diagram, name the type of force involved.*  **A.** Daya graviti *Gravitational force* **B.** Daya elastik *Elastic force* **C.** Daya apungan *Buoyant Force* **D.** Daya geseran *Frictional force*	Berdasarkan rajah berikut, namakan jenis daya yang dilabelkan dengan X. *Based on the following diagram, name the type of force labelled with X.*  **A.** Daya graviti *Gravitational force* **B.** Daya elastik *Elastic force* **C.** Daya geseran *Frictional force* **D.** Daya apungan *Buoyant Force* (*Uniformity of the arrangement of options' answers and additional label*)

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

8

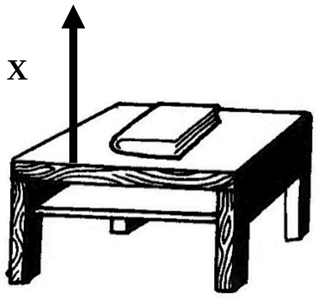
Rajah di bawah menunjukkan kedudukan sebuah buku.
The diagram below shows the position of a book.



Apakah daya tindak balas yang mewakili X?
What is the reaction force that represents X?

A. Daya graviti
Gravitational force
B. Daya normal
Normal force
C. Daya geseran
Frictional force
D. Daya apungan
Buoyant force

Rajah di bawah menunjukkan kedudukan sebuah buku.
The diagram below shows the position of a book.



Apakah daya tindak balas yang mewakili X?
What is the reaction force that represents X?

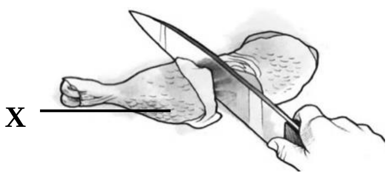
A. Daya graviti
Gravitational force
B. Daya normal
Normal force
C. Daya geseran
Frictional force
D. Daya apungan
Buoyant force

(*Changes in the label's position*)

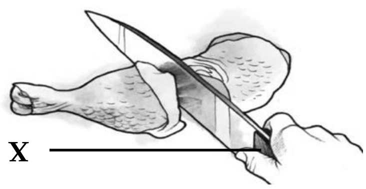
---

24 (a)

Puan Mala ingin memotong seekor ayam menjadi beberapa ketulan yang bersaiz kecil. Rajah di bawah menunjukkan peralatan X yang digunakan oleh Puan Mala.
Mrs. Mala wanted to cut a chicken into several small chunks. The diagram below shows the equipment X used by Mrs. Mala.



Puan Mala ingin memotong seekor ayam menjadi beberapa ketulan yang bersaiz kecil. Rajah di bawah menunjukkan peralatan X yang digunakan oleh Puan Mala.
Mrs. Mala wanted to cut a chicken into several small chunks. The diagram below shows the equipment X used by Mrs. Mala.



(*Changes in the label's position*)

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

After the improvements and modifications made to these items, all experts agreed that the improved instrument had content validity and was suitable to be administered in the actual studies. In addition, experts also agreed that all the items fit the six levels of the Revised Bloom's Taxonomy (Anderson & Krathwothl, 2001).

Reliability of FMAT Instruments

After conducting the pilot test, the reliability of the FMAT items was assessed using the Split-Half and Kuder-Richardson 20 (KR-20) methods. Table 5 shows the reliability value of FMAT instrument.

Table 5: Summary of Reliability analysis for FMAT Instrument

Section	Number of Item (N)	Coefficient Alpha
Section A	24	0.757
Section B	12	0.732

For items in Section A, the KR-20 method is more suitable to use because the items are in the dichotomous choices. Through data analysis, the coefficient alpha value for Section A is 0.757, which is above 0.70. It indicates that the Section A items of the FMAT instrument has an adequate reliability (Kaplan, & Saccuzzo, 2017). Furthermore, the reliability of the items in Section B is measured using the Split-Half method. The reliability analysis using the Split-Half method in this study produced an alpha coefficient value of 0.732, indicating that the items of Section B had adequate and acceptable reliability.

Item Difficulty Index

Researcher have conducted item analysis to explain the difficulty index (*p*) based on the results of the pilot test. The findings of the difficulty index (*p*) of all items in the FMAT instrument which consist of 24 items in Section A (objective), and 12 items in Section B (subjective) are shown in Table 6:

Table 6: Difficulty Index of Items in FMAT Instrument

Difficulty level	Items	Total	Recommendation
Easy	1, 2, 3, 4, 9, 10, 20, 23 (a), 25 (a)	9	Retained
Moderately difficult	5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21 (a), 21 (b), 22 (a), 22 (b), 23 (b), 23 (c), 24 (a), 24 (b), 25 (b)(i), 25 (c).	23	Retained
Difficult	24 (c), 24 (d), 25 (b)(ii), 25 (d)	4	Modified/removed

Based on the difficulty index, 9 (25%) of the items were easy, 23 (64%) items have the moderate level of difficulty, while the other 4 (11%) items were classified as difficult. Thus, the items that have been identified as difficult item should either be revised and improved or should be dropped from the FMAT instrument.

Discussion

An instrument that can be used to measure and determine the level of students' understanding of the Force and Motion concepts has been developed. The development of this FMAT instrument has adapted the format of the Final Academic Session Examination (UASA) for science subjects at the lower secondary level. The process of developing the FMAT instrument

Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

has gone through a systematic process and covers all the six level of Revised Bloom's Taxonomy.

In terms of validation, the content validity of the FMAT instrument was determined based on the Content Validation Index (CVI) obtained from three experts. As a result, three (3) items in the FMAT that had an i-CVI value of less than 1.00 must be revised and improved, such as changing the position of the labels, adding labels to the diagram, and ensuring the uniformity in the arrangement of the options' answers. On top of that, the 'S-CVI/average' value obtained for the FMAT instrument is 0.97, indicates that the FMAT instrument has a high degree of content validity. Besides that, the FMAT instrument also shows acceptable reliability with the coefficient alpha value in Section A is 0.757, and 0.732 for Section B. This means that the FMAT instrument is reliable and valid for use in actual studies.

Based on the item analysis, the difficulty index value of ($0.30>p>.070$) revealed that 23 out of the 36 items are 'good' (moderately difficult) and 9 items can be seen as 'fair' (easy). On the other hand, 4 items that failed to satisfy the condition are considered 'poor' items. Therefore, it is recommended to either removed or modified the items.

Conclusion

The focus of this study is to develop and evaluate the quality of the FMAT instrument that can be used in assessing students' conceptual understanding regarding Force and Motion topic. The study successfully achieved its objectives by producing an instrument that meets high validity and reliability standards, setting a benchmark for more precise and effective assessments of students' understanding of scientific concepts. A robust instrument like FMAT can significantly enhance educational outcomes by ensuring assessments accurately reflect students' knowledge and abilities. Besides, the FMAT instrument also enables teachers to identify gaps in students' understanding of the Force and Motion topic more effectively, providing actionable insights for tailoring instruction and interventions. This targeted approach can lead to improved learning outcomes in this fundamental area of the Science curriculum. The rigorous development process of the FMAT instrument emphasizes the importance of ensuring validity and reliability to minimize measurement errors and maximize the accuracy of the intended construct. This study contributes to the field of educational measurement by providing a framework and guidelines for teachers and researchers to develop high-quality achievement tests. Specifically, the study serves as a reference for teachers in deveoping instruments that align with curriculum objectives and assessment standards.

It is recommended that future studies incorporate advance item analysis techniques, such as item discrimination and distractor analysis, to further evaluate the effectiveness of test items. These analyses will help determine how well items differentiate between high and low achieving students, and assess the functionality of distractors in diverting students from the correct answer. This analysis will ultimately lead to better test quality that can assist teachers to improve teaching strategies and identify misconceptions among students. In addition, such refinements will contribute to developing even more reliable and valid instruments that enhance assessment practices and support data-driven instructional strategies.

In conclusion, this study highlights the significant role of the FMAT instrument in improving assessment practices and advancing educational outcomes. The successful implementation of the FMAT not only bridges gaps in assessing conceptual understanding but also serves as a

Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

blueprint for developing future assessment tools. This contribution to educational measurement and evaluation lays the groundwork for ongoing improvements in the design and application of assessment instruments across various domains.

Acknowledgments
The authors would like to acknowledge the Ministry of Education for providing financial support through the Hadiah Latihan Persekutuan (HLP) scholarship to the first author. This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors.

References

Abd Latif, A., Yusof, I. J., Mohd Amin, N. F., Libunao, W. H., & Yusri, S. S. (2016). Multiple-Choice Items Analysis Using Classical Test Theory And Rasch Measurement Model. *Man In India*, *96*(1-2), 173-181.

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy For Learning, Teaching And Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Badruldin, N. A., & Alias, S. N. (2022). Level of Conceptual Understanding Among Secondary Students on Topic Of Forces and Motion Using Half-Length Force Concept Inventory (HFCI). *ICCCM Journal of Social Sciences and Humanities*, *1*(2), 12-20. https://doi.org/10.53797/icccmjssh.v1i2.2.2022

Bahtaji, M. A. A. (2023). Examining the physics conceptions, science engagement and misconceptions of undergraduate students in STEM. *Journal of Baltic Science Education*, *22*(1), 10-19.

Basran, A., & Lajium, D. (2020). Aplikasi Model Rasch Dalam Pengujian Instrumen Inventori Konsep Daya. *International Journal of Modern Education*, *2*(6), 14-27. https://doi.org/10.35631/IJMOE.26003

Bekoe, S. O. (2023). Construction and Analysis of Items for Assessing Learning Outcomes: Concise Models for Ensuring Validity. *International Research in Education*, *11*(2), 29-51. https://doi.org/10.5296/ire.v11i2.21160

Bichi, A. A. (2015). Item Analysis Using A Derived Science Achievement Test Data. *International Journal Of Science And Research (IJSR)*, *4*(5), 1656-1662

Bichi, A. A., & Embong, R. (2018). Evaluating the Quality of Islamic Civilization and Asian Civilizations Examination Questions. *Asian People Journal*, *1*(1), 93-109.

Bouzid, T., Kaddari, F., & Darhmaoui, H. (2022). Force and motion misconceptions' pliability, the case of Moroccan high school students. *The Journal of educaTional research*, *115*(2), 122-132.

Ismail, A. T., & Ayop, S. K. (2016). Tahap Kefahaman Dan Salah Konsep Terhadap Konsep Daya dan Gerakan Dalam Kalangan Pelajar Tingkatan Empat. *Jurnal Fizik Malaysia*, *37*(1), 01090-01101.

Mohd Noor, A. (2021). Evaluating Multiple Choice Questions From Engineering Statistics Assessment. *International Journal Of Education And Pedagogy*, *3*(4), 33-46

Champlain, A. F. (2010). A Primer On Classical Test Theory And Item Response Theory For Assessments In Medical Education. *Medical Education*, *44*(1), 109-117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Che Isa, Z., & Azid, N. (2021). Analisis Statistik Kesahan dan Kebolehpercayaan Ujian Pencapaian Reka Bentuk Elektrik. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *6*(8), 196-206. https://doi.org/10.47405/mjssh.v6i8.904

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029
Chin, H., Thien, L. M., & Chiew, C. M. (2019). The Reforms of National Assessments In Malaysian Education System. *Journal of Nusantara Studies*, *4*(1), 93-111. https://doi.org/10.24200/Jonus.Vol4iss1pp93-111
Chua, Y. P. (2012). *Kaedah dan Statistik Penyelidikan: Kaedah Penyelidikan* (2nd Ed.). Mcgraw-Hill
Connor Desai, S., & Reimers, S. (2019). Comparing the use of open and closed questions for Web-based measures of the continued-influence effect. *Behavior Research Methods, 51,* 1426–1440. https://doi.org/10.3758/s13428-018-1066-z
Cooper, D. R., & Schindler, P. S. (2011). *Business Research Methods* (11th Ed.). Mcgraw-Hill.
Creswell, J. W. & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative and Mixed-Method Approaches* (5th Ed.). Sage Publications Ltd.
Davis, L. L. (1992). Instrument Review: Getting the Most from Your Panel of Experts. *Applied Nursing Research, 5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4
Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. (5th Ed.). Prentica-Hall Inc.
Esomonu, N. & Eleje, L. (2020). Effect of Diagnostic Testing on Students' Achievement in Secondary School Quantitative Economics. *World Journal of Education, 10*(3), 178-187. https://doi.org/10.5430/wje.v10n3p178.
Fraenkel, J.R & Wallen, N.E. (1996). *How to Design and Evaluate Research in Education* (3rd Ed.). Mc. Graw-Hill
Halim, L., & Meerah, T. S. M (2016). Science Education Research And Practice In Malaysia. In: MH. Chiu (Ed) *Science Education Research And Practice In Asia*. (pp. 71-93). Springer. https://doi.org/10.1007/978-981-10-0847-4_5
Hambleton, R. K., & Jones, R. W. (1993). Comparison *of* Classical Test Theory And Item Response Theory And Their Applications To Test Development. *Educational Measurement: Issues And Practice, 12*(3), 38-47.
Haw, L. H., & Sharif, S. (2020). The Development and Validation of Science Achievement Test. *Journal of Education And Practice, 11*(20). https://doi.org/10.7176/JEP/11-20-12
Haw, L. H., Sharif, S., & Han, C. G. K. (2022). Analyzing The Science Achievement Test: Perspective of Classical Test Theory And Rasch Analysis. *International Journal of Evaluation And Research In Education,* *11*(4), 1714-1724. https://doi.org/10.11591/Ijere.V11i4.22304
Henning, G. (1987). *A Guide To Language Testing: Development, Evaluation, Research*. Newberry House.
Hock, T. T., Ayub, A. F. M., Shah, M. M., & Ahamed, A. B. (2022). Implementation of Classroom-Based Assessment in Malaysia. *International Journal of Academic Research in Progress Education and Development*, *11*(4), 70-80. https://doi.org/10.6007/IJARPED/V11-I4/14621
Irwing, P., & Hughes, D. (2018). Test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. John Wiley & Sons Ltd.
Kamarrudin, H., Halim, L., & Mohtar, L. E. (2020). Sumbangan Penguasaan Asas Fizik dan Matematik Terhadap Keupayaan Pengaplikasian Fizik bagi Topik Mekanik Tingkatan Enam. *Jurnal Fizik Malaysia*, *41*(1), 10037-10050.
Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues.* Nelson Education

International Journal of Modern Education
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029
Khoirunnisa, R., Syuhendri, Kistiono, & Afifa, M. (2024). Misconceptions of High School Students on Motion and Force Using the Force Concept Inventory (FCI). *Jurnal Penelitian IPA, 10*(5), 2711-2720. https://doi.org/10.29303/jppipa.v10i5.6979

M. Yusop, S. R., Rasul, M. S., & Mohamad Yasin, R. (2024). Challenges, Strengths, and Relevance of Integrating Classroom Based Assessment in Technical and Vocational Education Training. *International Journal Of Evaluation And Research In Education, 13*(4), 2440-2449. https://doi.org/10.11591/Ijere.V13i4.27900

Magno, C. (2009). Demonstrating the Difference Between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University*, *13*(4), 310-315.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and Assessment in Teaching* (11th Ed.). Pearson Education Inc.

Mills, G. E., & Gay, L. R. (2016). *Educational Research: Competencies For Analysis And Application* (11th Ed.). Pearson.

Ministry of Education Malaysia (2016). *Dokumen Standard Kurikulum dan Pentaksiran Sains Tingkatan Dua*. Bahagian Pembangunan Kurikulum

Ministry of Education Malaysia (2018[a]). *Dokumen Standard Kurikulum dan Pentaksiran Sains Tingkatan Empat dan Lima.* Bahagian Pembangunan Kurikulum

Ministry of Education Malaysia (2018[b]). *Dokumen Standard Kurikulum dan Pentaksiran Fizik Tingkatan Empat dan Lima*. Bahagian Pembangunan Kurikulum

Ministry of Education Malaysia. (2020[a]). *Sijil Pelajaran Malaysia: Format Pentaksiran Mulai Tahun 2021*. Lembaga Peperiksaan. Ministry of Education Malaysia. (2020[b]). *Laporan Kebangsaan TIMSS 2019 : Trends in International Mathematics and Science Study*. Bahagian Perancangan dan Penyelidikan Dasar Pendidikan.

Ministry of Education Malaysia (2022). *Buku Format Instrumen Pentaksiran dan Pelaporan Ujian Akhir Sesi Akademik Sekolah Menengah*. Lembaga Peperiksaan.

Ministry of Education Malaysia. (2023). *Laporan Awal Pencapaian Malaysia dalam Programme for International Student Assessment (PISA) 2022.* Bahagian Perancangan dan Penyelidikan Dasar Pendidikan.

Murshed, M., Phang, F. A., Bunyamin, M. A. H. B., & Yusof, I. J. (2020). The reliability analysis for force concept inventory. *International Journal of Psychosocial Rehabilitation, 24*(05), 143-151. https://doi.org/10.37200/IJPR/V2415/PR201677

Narinasamy & Nor 'Aidah Nordin (2018). Implementing Classroom Assessment in Malaysia: An Investigation, *Jurnal Kurikulum*, 3, 55–63. https://doi.org/10.6007/IJARPED/V9-I2/7499

Nugba, R. M., & Quansah, F. (2021). Standardized Achievement Testing, Aptitude Testing, and Attitude Testing: How Similar or Different Are These Concepts in Educational Assessment. *Asian Journal Of Education And Social Studies, 15*(3), 42-54. https://doi.org/10.9734/AJESS/2021/v15i330383

Obilor, E. I., & Omeke, K. S. (2024). Use of Test Blueprint in Improving Teachers' Test Construction Skills for Quality Assessment. *Journal of Educational Research and Development*, *11*, 127-139.

Ong, E. T, & Mohamad, M. A. J. (2014). Pembinaan Dan Penentusahan Instrumen Kemahiran Proses Sains untuk Sekolah Menengah. *Jurnal Teknologi*, *66*(1), 7-20.

Volume 6 Issue 23 (December 2024) PP. 421-438
DOI: 10.35631/IJMOE.623029

Phang, F. A., Khamis, N., Nawi, N. D., & Jaysuman Pusppanathan. (2021). TIMSS 2019 Science Grade 8: Where is Malaysia standing?. *Asean Journal of Engineering Education*, *4*(2). https://doi.org/10.11113/ajee2020.4n2.10

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*(3), 489-497. https://doi.org/10.1002/Nur.20147

Reynolds, C. R., Altmann, R. A., Allen, D.N. (2021). How to Develop a Psychological Test: A Practical Approach. In *Mastering Modern Psychological Testing* (pp. 663–692). Springer, Cham. https://doi.org/10.1007/978-3-030-59455-8_18

Shahari, S., & Phang, F. A. (2023). Tahap Kefahaman Konsep Daya Dan Kerangka Alternatif Pelajar Matrikulasi Melalui Ujian Penilaian Konsep Daya. *Jurnal Pendidikan Bitara UPSI*, *16*(1), 54-66. https://doi.org/10.37134/bitara.vol16.1.6.2023%20

Suppiah Shanmugam, S. K., Wong, V ., & Rajoo, M. (2020). Examining the Quality of English Test Items Using Psychometric and Linguistic Characteristics Among Grade Six Pupils. *Malaysian Journal Of Learning And Instruction*, *17*(2), 63-101. https://doi.org/10.32890/Mjli2020.17.2.3

Volfson, A., Eshach, H., & Ben-Abu, Y. (2020). Identifying physics misconceptions at the circus: The case of circular motion. *Physical Review Physics Education Research, 16*(1), 1-11. https://doi.org/10.1103/PhysRevPhysEducRes.16.010134

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education In Medicine Journal*. *11*(2), 49–54. https://doi.org/10.21315/Eimj2019.11.2.6