Jurnal Materi dan Pembelajaran Fisika
Volume 15, Number 2, pp. 86-91 2025
Website : https://jurnal.uns.ac.id/jmpf

# Development of Scientific Comprehension Assessment Instruments With Two-Tier Multiple Choice Assisted by Thunkable on the Topic of Work and Energy

**Ravena Nawang Ara*, Sukarmin**
*Physics Education Study Program, Faculty of Teacher Training and Education, Sebelas Maret University, Ir. Sutami No.36A Street, Jebres, Surakarta, Central Java, 57126, Indonesia*
*Corresponding Author Email : nevaranawangara@gmail.com

**Article's Info**
Received: 6th, May, 2025
Accepted: 26th, November, 2025
Published: 30th, November, 2025

**DOI**:
https://doi.org/10.20961/jmpf.v15i2.101941

**How to Cite :** Ara, R. N., & Sukarmin, S. (2025). Pengembangan Instrumen Penilaian Pemahaman Sains dengan *Two-Tier Multiple Choice* Berbantuan Thunkable pada Materi Usaha dan Energi. *Jurnal Materi dan Pembelajaran Fisika,* 15(1), 86-91

***Abstract***. This study aims to develop a instrument for assessing scientific comprehension with Two-Tier Multiple Choice Assisted by Thunkable on the Topic of Work and Energy that meets the criteria as a good assessment instrument. This study uses a development research method from the Wilson, Oriondo, and Antonio (1998) model, namely the test design stage which begins with determining the purpose of the test, determining the competencies tested, determining the material tested, preparing the test grid, writing items, validating items, improving items and assembling tests, and preparing scoring guidelines, then instrument validation carried out by determining the test subjects, conducting trials, and analyzing the data from the trial results and completing with a broad test measurement. The data collection method used was a questionnaire using a two-tier multiple choice instrument in the form of a thunkable-assisted application. The data sources were 6 experts and 270 students from 3 schools, namely SMA Negeri 1 Sukoharjo, SMA Negeri 1 Tawangsari, and SMA Negeri 1 Bulu. The results showed that of the 22 items of scientific comprehension assessment instruments with two-tier multiple choice assisted by thunkable on the topic of work and energy with the Quest program, 20 items were declared fit with the Rasch model. The conclusion of this research is that the instrument of scientific comprehension with two-tier multiple choice assisted by thunkable on the topic of work and energy as a good assessment instrument based on the Rasch model.

**Keywords:** Assessment Instrument, Two-Tier Multiple Choice, Scientific Comprehension, Thunkable.



## INTRODUCTION

Education plays a very important role in the development of a country that functions to form a smart generation, character, and ready to face various global challenge (A et al., 2023). So that, not only acts as a tool to channel knowledge, but also as a tool to create a quality individual personality and character. As an asset for every individual, through the education process, a person can explore their latent potential. In addition, participation can improve the quality of an individual (Cahyani et al., 2020). Furthermore obtained through the learning process at school is expected to create graduates who can apply their knowledge in solving everyday problems. Besides that, education also acts as an important foundation for everyone to think critically and creatively (Fahmi et al., 2023). In line with this strategic role, curriculum renewal is important to ensure that education remains relevant to social and technological changes.

In line with the development of the Industrial Revolution 4.0, the Merdeka Curriculum was introduced as a response to the competitive challenges of the 21st century. This curriculum is designed to create meaningful learning while preparing the younger generation to face the changing times (Indarta et al., 2022). Its flexibility allows educators and schools to adapt the curriculum to social developments, student needs, and advancements in science and technology (Akhmadi, 2023). The emphasis on scientific literacy in this curriculum is relevant to the demands of the digital era, where young people are required to be technologically literate, think critically, and collaborate to solve complex problems. The differentiated approach in the Merdeka Curriculum also facilitates science learning tailored to students' interests and talents, thereby fostering intrinsic motivation and creativity.

Preliminary interviews at SMAN 1 Sukoharjo revealed that learning assessments have so far been conducted manually using printed instruments, without utilizing digital applications. 80% of teachers have also never used two-tier multiple-choice instruments, relying only on conventional multiple-choice questions and essays due to limited knowledge of its structure and scoring technique as well as the lack of available digital tools that support its implementation. This situation indicates the need for developing more accurate and practical assessment instruments. Amid the rapid advancement of educational technology, digital-based instruments accessible via mobile devices present a potential solution. This approach not only enhances assessment efficiency (such as score input and reduced printing costs) but also has the potential to increase student engagement in the evaluation process.

The two-tier multiple-choice instrument is an effective tool for measuring in-depth scientific understanding, as it not only evaluates their conceptual knowledge but also probes the underlying reasoning behind their answers, thereby enabling the identification of scientific misconceptions more accurately (Chandrasegaran et al., 2007) This instrument consists of two levels of questions: the first tier tests problem identification and analysis skills, while the second tier evaluates the reasoning behind the chosen answers. This complexity minimizes guesswork and encourages students to reflect on their understanding. The scoring guidelines in the form of two-tier multiple choice are based on the scoring guidelines of Bayrak (2013) which allow a detailed categorization of students' conceptual mastery and misconception patterns. In a digital context, the development of such instruments is further facilitated by platforms like Thunkable, which enables the creation of interactive assessment applications with features such as randomized questions, automatic score settings, and answer locking (Fauzi, 2020). With this approach, users can design applications without the need to write code manually (Anam et al., 2022). These advantages make Thunkable a promising tool for creating objective, efficient, and engaging evaluations (Adiarta & Divayana, 2019).

This research focuses on the topic of work and energy, which includes four subtopics: the concept of work, kinetic energy, potential energy, and the relationship between them. This material was chosen due to its applicability in daily life and its potential to be developed into a challenging two-tier multiple-choice instrument. Research also shows that many students still experience misconceptions related to work–energy principles, especially in differentiating force and work, as well as interpreting changes in mechanical energy (Halilović et al., 2022). Understanding these concepts trains students to connect physics variables with real-world phenomena. Examples of relevant real-world contexts include energy conversion in roller coasters, the lifting of objects against gravity, and changes in kinetic energy when a vehicle accelerates or brakes.

Several previous studies have developed similar instruments but with various limitations. (Mufti & Sunarti (2024) used a five-tier diagnostic test, which, although comprehensive, required considerable time for development and data analysis. Wati (2024) developed a four-tier test for identifying misconceptions, but the instrument was underperforming for large-scale implementation. Meanwhile, Pangesti (2024) implemented a three-tier multiple-choice test that was more accurate than the two-tier version, but the complexity of scoring posed a challenge. Pratiwi et al (2024) utilized Google Forms for a two-tier test, but the platform was deemed less interactive and had design limitations.

To address the gaps identified above, this study proposes the development of assessment instrument use two-tier multiple choice by the title "Development of Scientific Comprehension

Assessment Instruments With Two-Tier Multiple Choice Assisted by Thunkable on the Topic of Work and Energy".

## METHOD

The research conducted is developmental research, commonly known as Research and Development (R&D). The study follows the development model proposed by Wilson, Oriondo, and Antonio (1998) as cited in (Istiyono, 2014), which encompasses several key stages. The process begins with the test design phase, which includes determining the test objectives, identifying the competencies to be assessed, selecting the relevant subject matter, creating test blueprints, writing test items, conducting item validation, revising items and assembling the final test, as well as developing scoring guidelines. Subsequently, is the test trial phase, involving the selection of trial subjects, implementation of the trial, and analysis of the trial data. The final phase consists of conducting large-scale measurement tests (Istiyono, 2014).

The quantitative assessment from the validators was analyzed using Aiken's V formula (Aiken, 1985). The evaluation process employing this formula involves scoring according to established categories, with V values ranging between 0 and 1. Content validity can be determined using Aiken's V formula (Aiken, 1985), as follows:

$$V = \frac{\Sigma S}{n(c-1)}; S = r - L_o \quad [1]$$

where:
$V =$ Aiken's validity index
$S =$ Sum of validator ratings minus the lowest possible rating value
$L_o =$ Lowest possible rating value (minimum scale value)
$c =$ Highest possible rating value (maximum scale value)
$r =$ Rating score given by a validator
$n =$ Number of validator

The results obtained are categorized as table 1.

**Table 1**. Interpretation Range for Validity Scores

| Category | Validity Scores |
|---|---|
| Very High | $0{,}80 - 1{,}00$ |
| High | $0{,}60 - 0{,}79$ |
| Medium | $0{,}40 - 0{,}59$ |
| Low | $0{,}20 - 0{,}39$ |
| Very Low | $< 0{,}20$ |

(Kusumasari et al., 2025)

The instrument was then tested on a limited basis (initial trial) to 3 different schools, namely SMA Negeri 1 Sukoharjo, SMA Negeri 1 Tawangsari, and SMA Negeri 1 Bulu. The results of the scientific comprehension assessment instrument were analyzed with the Quest program to determine the estimated item difficulty index, the estimated item fit with the Rasch model, and the estimated item acceptability.

## RESULT AND DISCUSSION

As part of the content validity evaluation process, the validation using Aiken's V formula aims to systematically gather expert judgments because it is suitable for analyzing ordinal rating scales provided from six specialists, evaluating three key aspects: indicators, construction, and language. The questionnaire results were analyzed using Aiken's V formula, and the result as shown in Table 2.

**Table 2.** Validity Test Results with Aiken's V

| No. | V | Category |
|---|---|---|
| 1 | 1 | Very High |
| 2 | 0,94 | Very High |
| 3 | 1 | Very High |
| 4 | 0,88 | Very High |
| 5 | 0,94 | Very High |
| 6 | 0,88 | Very High |
| 7 | 0,77 | High |
| 8 | 0,88 | Very High |
| 9 | 0,88 | Very High |
| 10 | 1 | Very High |
| 11 | 1 | Very High |
| 12 | 0,83 | Very High |
| 13 | 0,77 | High |
| 14 | 0,88 | Very High |
| 15 | 0,94 | Very High |
| 16 | 0,94 | Very High |
| 17 | 0,94 | Very High |
| 18 | 0,88 | Very High |
| 19 | 0,77 | High |
| 20 | 1 | Very High |
| 21 | 0,94 | Very High |
| 22 | 0,88 | Very High |

The instrument validation stage was carried out by 6 experts. The results of the acquisition and calculation of the validation of the assessment instrument found that all items obtained an Aiken's V value of more than 0.75 so that these items had very good content validity (Azwar, 2012). In detail, 19 items entered the very high category, and 3 other items entered the high category, where all items had a high level of relevance to the measurement objectives. In the pilot and broad test stages, the analysis was conducted using the QUEST program with the form of polytomic data. Based on the Rasch model analysis shows that the INFIT MNSQ value of all items ranges from 0.77 - 1.33 so that it fits the Rasch model. This value range indicates that the responses to the items fit the Rasch model assumptions, as INFIT MNSQ values between 0.77–1.30/1.33 are recommended for acceptable item fit (Bond & Fox, 2015). INFIT MNSQ is used to evaluate the internal consistency of item responses with greater sensitivity to unexpected responses near a person's ability level, thereby indicating whether the item measures the intended construct.

In addition, based on the threshold value or the level of difficulty, it is found that there are 3 items in the difficult category or 15% and 17 items in the moderate category or 85%. The fit component on the OUTFIT t value of items 1 and 2 failed while items 3 to 20 passed indicating that the passed items could be used and obtained a person reliability of case estimate value of 0.76 with a moderate category. OUTFIT statistics detect unexpected responses by individuals far from the item's difficulty level, which is important to ensure that each item functions consistently across all student ability levels (Boone et al., 2014).

The results of the QUEST-based Rasch model analysis indicate that the two-tier multiple-choice items have met the requirements for a good assessment instrument in terms of item fit, difficulty distribution, and reliability. Based on the stages that have been carried out entirely, the final product of this research is a science understanding assessment instrument with two-tier multiple choice assisted by thunkable on the material of effort and energy. The results of the analysis obtained using the QUEST program and based on the Rasch model, the scientific comprehension assessment instrument with two-tier multiple choice assisted by thunkable on the topic of work and energy meets the criteria as a good assessment instrument.

## CONCLUSION

Based on the results of the research that has been carried out, it can be concluded that the study successfully developed a science comprehension assessment instrument using two-tier multiple choice questions through the Thunkable platform, focusing on work and energy concepts. Following the Wilson, Oriondo, and Antonio (1998) development model, the instrument was systematically created through test design, construction, and implementation phases. It comprehensively covers four key physics concepts: mechanical work, kinetic energy, potential energy, and their interconversion, ensuring thorough content coverage aligned with learning objectives. Validation results confirmed the instrument's strong psychometric properties, with all items scoring above 0.75 in Aiken's V analysis. 19 items achieved "excellent" validity status while 3 items received "high" validity ratings, demonstrating robust content validity. The Rasch model analysis further verified the instrument's quality, showing all 20 items had appropriate difficulty levels (15% challenging, 85% moderate) and good statistical fit (INFIT MNSQ range: 0.77-1.33), with satisfactory reliability (0.76 case estimate). This digitally enhanced assessment tool effectively combines rigorous measurement standards with practical classroom applicability through its innovative two-tier question format and Thunkable platform integration. The instrument provides educators with a reliable means to evaluate students' conceptual understanding of work and energy principles while demonstrating the successful application of research and development models in educational technology. Its development process serves as a valuable template for creating similar technology-based assessment instruments in other science domains. Additionally, future development could integrate enhanced digital features such as student response analytics, instant feedback, and scalability across different physics topics to maximize its practicality in classroom assessment.

## REFERENCES

A, N., Norrahman, R. A., Muhammadong, & Wibowo, T. S. (2023). Pemberdayaan Sumber Daya Manusia dalam Konteks Manajemen Pendidikan. *Journal Of International Multidisciplinary Research*, *1*(2), 222–235.

Adiarta, A., & Divayana, D. G. H. (2019). Pengembangan Soal Test Digital Matakuliah Asesmen Dan Evaluasi Menggunakan Aplikasi Wondershare. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *16*(2), 287. https://doi.org/10.23887/jptk-undiksha.v16i2.19199

Akhmadi, A. (2023). Implementasi Kurikulum Merdeka di Madrasah Ibtidaiyah. *Andragogi: Jurnal Diklat Teknis Pendidikan Dan Keagamaan*, *11*(1), 33–44. https://doi.org/10.36052/andragogi.v11i1.310

Anam, M. K., Abbas, D. S., & Anggraini, L. (2022). Meningkatkan Literasi Perbankan Syariah dengan Mengembangkan Aplikasi Edukasi Berbasis Android. *Sci-Tech Journal*, *2*(2), 96–104. https://doi.org/10.56709/stj.v2i2.69

Bayrak, B. K. (2013). Using Two-Tier Test to Identify Primary Students' Conceptual Understanding and Alternative Conceptions in Acid Base. *Mevlana International Journal of Education*, *3*(2), 19–26. https://doi.org/10.13054/mije.13.21.3.2

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Springer. https://doi.org/10.1007/978-94-007-6857-4

Cahyani, A., Listiana, I. D., & Larasati, S. P. D. (2020). Motivasi Belajar Siswa SMA pada Pembelajaran Daring di Masa Pandemi Covid-19. *IQ (Ilmu Al-Qur'an): Jurnal Pendidikan Islam*,

*3*(01), 123–140. https://doi.org/10.37542/iq.v3i01.57

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. Chemistry Education Research and Practice, 8(3), 293 – 307.

Fahmi, J., Nahadi, N., & Hernani, H. (2023). Pengembangan Asesmen Formatif Berbasis Problem Based Learning untuk Meningkatkan Keterampilan Berpikir Kritis: Need Assessment Study. *Orbital: Jurnal Pendidikan Kimia*, *7*(2), 237–249. https://doi.org/10.19109/ojpk.v7i2.19922

Fauzi, I. M. D. (2020). *Pengembangan media mobile learning menggunakan thunkable pada materi SPLTV*.

Halilović, A., Mešić, V., & Hasović, E. (2022). *The post-instruction conceptions about conservation of mechanical energy: Findings from a survey research with high school and university students.* Journal of Turkish Science Education, 19(1), 144-162. DOI: 10.36681/tused.2022.115

Indarta, Y., Jalinus, N., Waskito, W., Samala, A. D., Riyanda, A. R., & Adi, N. H. (2022). Relevansi Kurikulum Merdeka Belajar dengan Model Pembelajaran Abad 21 dalam Perkembangan Era Society 5.0. *Edukatif : Jurnal Ilmu Pendidikan*, *4*(2), 3011–3024. https://doi.org/10.31004/edukatif.v4i2.2589

Istiyono, E. (2014). DEVELOPING HIGHER ORDER THINKING SKILL TEST OF PHYSICS (PhysTHOTS) FOR SENIOR HIGH SCHOOL STUDENTS. *Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. http://journal.uny.ac.id/index.php/jpep/article/view/2120

Kusumasari, E. D., Muhtarom, & Sumarno. (2025). Jurnal Pengembangan dan Penelitian Pendidikan. *Jurnal Pengembangan Dan Penelitian Pendidikan*, *07*(1), 154–173.

Mufti, M. B., & Sunarti, T. (2024). *Identifikasi Miskonsepsi Siswa Materi Usaha dan Energi Menggunakan Five Tier Diagnostic Test*. *13*(3), 191–200.

Pangesti, D. I. (2024). *Unnes Physics Education Journal Analisis Keterampilan Berpikir Tingkat Tinggi Siswa SMA Menggunakan Instrumen*. *13*(1), 87–96.

Pratiwi, I. T., Viyanti, & Permadi, D. (2024). The Development of the Two-Tier Diagnostic Test Instrument with Google Form to Measure Student Misconceptions on Energy and Energy Forms. *Impulse: Journal of Research and Innovation in Physics Education*, *4*(1), 22–30. https://doi.org/10.14421/impulse.2024.41-03

Wati, W. (2024). *Pemetaan Miskonsepsi Mahasiswa Fisika pada Konsep Energi , Kinematika , dan Listrik Statis dengan Tes Disnogtik Four-Tier*. *4*, 808–820.