

Adaptation and Validation of a Four-Tier Misconception Diagnostic Test on Work and Energy Using Rasch Modeling

Gress Faihah, Moh. Irma Sukarelawan*, and Toni Kus Indratno

Physics Education Study Program, Faculty of Teacher Training and Education,
Universitas Ahmad Dahlan, Yogyakarta, Indonesia

*irma.sukarelawan@pfis.uad.ac.id

Article Info	Abstract
Article history: Received: September 13, 2025 Accepted: November 2, 2025 Published: November 5, 2025 Keywords: Cross-cultural adaptation Four-tier test I-WPECT Instrument validation Rasch model This article is licensed under a creative commons attribution-share alike 4.0 international license © 2025 Jurnal ilmiah Pendidikan Fisika	The WPECT instrument has been globally tested for its quality in diagnosing student misconceptions. However, this instrument cannot be directly used in the Indonesian cultural context due to language differences. Therefore, a cross-cultural adaptation process is required. This study aims to adapt and cross-culturally validate a four-tier diagnostic test on the concepts of Work, Power, and Energy using the Rasch model (1PL) based on item difficulty estimation. The research was conducted using a quantitative survey method involving 125 students from SMA Negeri 1 Jetis. The focus of the study is to adapt and validate the 4-tier Work, Power, and Energy instrument to the Indonesian context. The analysis results show that the item reliability index is 0.87 and the Cronbach's Alpha is 0.81, indicating a high level of internal consistency within the instrument. The distribution of item difficulty is spread across four categories, with the majority of items classified as difficult. Twelve out of thirteen items meet the Rasch model fit criteria, while one item can still be retained as its outfit MNSQ value remains below the acceptable threshold (< 2.0). The instrument also fulfills the assumption of unidimensionality, as all 13 items measure a single underlying construct—namely, the concept of work, power, and energy. Furthermore, no gender-based item bias was detected. Therefore, the four-tier diagnostic test on the topic of work and energy is suitable for identifying misconceptions among high school students in Indonesia, validly and equitably. The availability of this adapted instrument can help educators in Indonesia more accurately diagnose student misconceptions and design more targeted instructional interventions.
To cite this article: Faihah, G., Sukarelawan, M. I., & Indratno, T. K. (2025). Adaptation and validation of a four-tier misconception diagnostic test on work and energy using rasch modeling. <i>Jurnal Ilmiah Pendidikan Fisika</i>, 9 (3), 436-447.	

INTRODUCTION

Education is one of the main pillars of human resource development. To achieve effective learning objectives, a deep understanding of concepts is very important. However, in the process, it is often found that students experience misconceptions. Misconceptions are misunderstandings or understandings that are not in accordance with scientific concepts that have been agreed upon by experts (Sukarelawan et al., 2021; Suprpto, 2020). This can occur due to misunderstandings between one concept and another, causing students to have incorrect views (Mbonvirivuze et al., 2019). Misconceptions can hinder the learning process and make it difficult for students to understand more complex material. Therefore, a high-quality instrument is an important step in accurately identifying and diagnosing misconceptions.

Misconceptions occur in almost all subjects, such as biology ([Fuchs et al., 2021](#)), mathematics ([Sari et al., 2024](#)), chemistry ([Härmälä-Braskén et al., 2020](#)), and physics ([Sukarelawan, Indratno, et al., 2024](#); [Sukarelawan, Sa'adah, et al., 2024](#)). Specifically in physics, misconceptions have a significant impact because understanding one concept often forms the basis for understanding subsequent concepts. If students have an incorrect understanding at the beginning, they will have difficulty when learning more complex material later on ([Ravish, 2025](#); [Syuhendri, 2022](#)).

Several studies indicate that students in Indonesia continue to experience physics misconceptions. A study in Padang found that, on average, students experienced misconceptions on the subtopics of the application of Newton's First, Second, and Third Laws, with percentages of 84.6%, 76.3%, and 85% ([Bawamenewi, 2024](#)). In the Jambi region, only 7.7% of junior high school students truly understood the concept of light ([Maison et al., 2021](#)). Not only students, but also university students can have misconceptions about temperature and heat. More than 60% of physics education students hold misconceptions on the subtopic of thermal equilibrium ([Fenditasari et al., 2020](#); [Sukarelawan et al., 2019](#)). Additionally, 79.45% of prospective physics teachers believe that heavier objects fall faster, and 47.95% hold misconceptions that gravity increases when objects fall ([Syuhendri, 2017, 2019](#)).

In an effort to identify misconceptions more accurately, experts are competing to develop various forms of diagnostic test instruments, ranging from conventional multiple-choice ([Bentaleb et al., 2022](#)) to Level 2 multiple-choice ([NASIR et al., 2023](#)) and Level 3 ([Irmak et al., 2023](#)) tests, and ultimately, Level 4. To date, the four-tier diagnostic test instrument is recognized as more accurate than previous versions. The four-tier diagnostic test instrument consists of four levels of questions: answer choices, reasons for the choices, confidence in the answers, and confidence in the reasons. This instrument not only assesses the correctness of answers but also explores the extent to which students believe the concepts they consider true. Thus, the four-tier instrument provides a more comprehensive picture of student understanding.

On the other hand, the procedure for developing diagnostic test instruments is not a short and simple process, but rather requires high expertise and ability, as well as a lengthy process. However, this is not an obstacle, as teachers or practitioners can use instruments that experts have developed in various countries through a cultural adaptation procedure. This procedure allows the application of instruments that have been proven to be of high quality through a more concise and simple procedure, without compromising quality.

One of the diagnostic misconception test instruments that has been globally recognized for its quality in the subject of work, power, and energy is the WPECT instrument developed by [Irmak et al. \(2023\)](#). This instrument has undergone a long and rigorous development process and has been empirically validated on 709 respondents. However, this instrument is still in English, so it cannot be directly applied to the context of Indonesian students. In addition, WPECT is still in a three-tier format. In the Indonesian context, Anggraini has conducted empirical validation of a four-tier diagnostic test instrument for work and energy materials on 50 11th-grade students ([Anggrayni & Ermawati, 2019](#)). This instrument does not encompass the concept of power, and the sample size is limited to 50 students. In the Indonesian senior high school physics curriculum, the concept of power is explicitly included as part of the core competencies and learning standards for the Work and Energy topic. Hence, the absence of the power concept in previous instruments indicates a gap that needs to be addressed to ensure alignment with the national curriculum. In Rasch modeling, a sample size of 50 only achieves a stability level of 1 logit with a 95% confidence level. In line with this, Aripiani developed a diagnostic test called FORTUNE ([Aripiani et al., 2023](#)). The material

developed in FORTUNE is limited to the concepts of work and energy and involves 70 students for empirical validation. In Rasch modeling, a sample size of 70 only achieves a stability level of $\frac{1}{2}$ logit with a 95% confidence level. To achieve a 99% confidence level at a stability level of $\frac{1}{2}$ logit, a sample size in the range of 108-243 is required ([Linacre, 1994](#)). Therefore, in this study, a sample size of 125 students was selected to ensure sufficient stability and precision in parameter estimation based on the Rasch model.

Although diagnostic test instruments for work and energy have been widely developed, there are still gaps found, namely that the Indonesian context does not yet represent the concept of power, and the sample size is limited. Therefore, this study aims to fill the existing research gap by developing diagnostic test instruments for work and energy by involving the concept of power and using a representative sample size to obtain a higher level of stability and confidence. The topic of work, power, and energy was chosen because these are fundamental yet conceptually challenging topics in physics education. Moreover, in the Indonesian senior high school curriculum, “power” is explicitly stated as part of the core competencies within the Work and Energy topic, highlighting its educational relevance.

The Rasch model was employed in this study because it provides several methodological advantages over classical test theory. Specifically, Rasch modeling enables the estimation of both item difficulty and person ability on a common logit scale, providing objective measurement that is independent of the sample and test form. It also provides fit statistics that allow empirical validation of each test item's quality and ensure that the measurement model conforms to the expectations of unidimensionality and construct validity. Thus, the use of the Rasch model strengthens the rigor and interpretability of the instrument validation process.

In the context of Indonesian education, students often have difficulty understanding abstract physics concepts due to differences in curriculum types, cultural contexts, and linguistic nuances when compared to the country of origin, where the WPECT instrument was developed. Therefore, the direct use of English-language instruments is unable to identify students' misconceptions accurately. Contextual adaptation is important so that the instruments developed are truly valid, relevant, and fair for use by students in Indonesia.

METHODS

This study is a quantitative study using a survey approach. The focus of the study is to adapt and validate the 4-tier work, power, and energy instrument to the Indonesian context. The subjects in this study were students of SMA Negeri 1 Jetis, consisting of two 11th-grade classes and two 12th-grade classes. There were a total of 130 students, consisting of 28 male students and 102 female students. The sampling technique used was simple random class sampling. Of this number, 125 respondents were included in the data analysis, as the other five respondents were excluded because they had extreme values that could have affected the validity of the analysis results. The extreme values used are based on the output from the Winsteps analysis. This sample size meets the stability level of up to $\frac{1}{2}$ logit with a confidence level of 99% ([Linacre, 1994](#)).

The main instrument used in this study was the 4-tier I-WPECT (Indonesian-Work Power Energy Concept Test) diagnostic test, which was adapted from the English version of the WPECT (Work Power Energy Concept Test) instrument developed by [Irmak et al. \(2023\)](#). This instrument consists of 13 questions that measure students' misconceptions on the topics of work, power, and energy. Each question in the I-WPECT consists of four levels, namely:

Tier 1: Conceptual questions

Tier 2: Belief in the answers at the first Tier

Tier 3: Reasons for the answers at the first Tier

Tier 4: Belief in the reasons chosen

Before use, the instrument was translated into Indonesian and adapted to the local cultural context (Lv et al., 2024; Sukarelawan, Indratno, et al., 2024). The linguist carried out the validation process to assess readability, clarity of terms, and contextual appropriateness. Revisions were made based on suggestions provided by the validator.

Data analysis was performed using Rasch modeling with the assistance of Winsteps software (Linacre, 2021). The data analyzed included (1) item and person statistics, (2) the Item Wright Map, (3) model fit analysis, (4) unidimensionality, and (5) differential item functioning analysis based on gender. Item and person statistics were used to show the quality of the items and the characteristics of the respondents. The Item Wright Map was used to see the distribution of item difficulty and student ability. Model fit analysis (fit statistics) was determined based on the infit and outfit mean square (MNSQ) values, which ideally should be in the range of 0.5 to 1.5. The unidimensionality test is satisfied if the variance explained by the measure is $\geq 30\%$ and the unexplained variance in the 1st contrast is $\leq 15\%$. Differential Item Functioning (DIF) analysis is reviewed based on gender. DIF analysis uses the DIF contrast value, with the neutral category ranging from -0.5 to 0.5 (no gender bias).

The research began with the translation and adaptation of the instrument from the original version to Indonesian. After validation by experts, the instrument was compiled in Google Form format. Next, the researchers obtained permission from schools and coordinated the data collection schedule. Data collection was conducted offline over a period of 1 to 2 weeks. Each student was given 30–40 minutes to complete the 13 questions in the instrument.

RESULTS AND DISCUSSION

Sample of Instrument

The following is a sample display of the instruments developed, as shown in Table 1.

Table 1 Sample display of the instruments developed

Original item (3-Tier)	Indonesian version (4-Tier)
As shown in the figure given below, the weightlifter lifts the barbell from the ground first (first case) and then holds it for 10 seconds above (second case).   First case Second case	Seperti yang ditunjukkan pada gambar di bawah ini, seorang atlet angkat besi mengangkat barbel dari tanah (kasus pertama), lalu menahannya selama 10 detik di atas (kasus kedua).   Kasus Pertama Kasus Kedua
1.1. Which of the following statements is correct about the work done on the barbell by the weightlifter in both cases? - Work is done only in the first case.* - Work is done only in the second case. - No work is done in both cases - Work is done in both cases	1.1 Manakah dari pernyataan berikut yang benar mengenai usaha yang dilakukan atlet angkat besi pada barbel di kedua kasus tersebut? - Usaha hanya dilakukan pada kasus pertama.* - Usaha hanya dilakukan pada kasus kedua.

Original item (3-Tier)	Indonesian version (4-Tier)
	c. Tidak ada usaha yang dilakukan pada kedua kasus. d. Usaha dilakukan pada kedua kasus.
1.2. What is the reason for your answer given above?	1.2 Apakah Anda yakin dengan jawaban Anda pada pertanyaan 1.1?
a. The work is done because energy is consumed. b. The work is done because a force is acting in the direction of motion.* c. Work is done whenever there is a motion. d. No work is done because there is no displacement in the vertical direction. e. The force acting on an object does work even if the object does not move. f. Other opinions. Please write: ...	a. Sangat tidak yakin b. Tidak yakin c. Cukup yakin d. Yakin e. Sangat yakin
1.3. Are you sure about your answers to the previous two questions?	1.3 Apa alasan jawaban Anda pada pertanyaan 1.1?
a. Sure b. Not sure	a. Usaha dilakukan karena energi dikonsumsi. b. Usaha dilakukan karena ada gaya yang bekerja searah dengan arah gerak. * c. Usaha dilakukan setiap kali ada gerakan. d. Tidak ada usaha yang dilakukan karena tidak ada perpindahan dalam arah vertikal. e. Gaya yang bekerja pada suatu benda tetap melakukan usaha meskipun benda tersebut tidak bergerak.
	1.4 Apakah Anda yakin dengan jawaban Anda pada pertanyaan 1.3?
	a. Sangat tidak yakin b. Tidak yakin c. Cukup yakin d. Yakin e. Sangat yakin

I-WPECT Statistical Summary

The statistical summary of the items in the I-WPECT instrument is summarized in Table 2.

Table 2 I-WPECT statistical summary

Index	Person	Item
Infit MnSq	1.10	1.01
Infit ZStd	0.04	0.04
Outfit MnSq	1.08	1.08
Outfit ZStd	0.02	0.45
Separation	1.69	2.64
Reliability	0.74	0.87
Mean	0.10	0.00
Standard deviation	0.60	0.23
Cronbach Alpha (KR-20)	0.81	

Based on Table 2, a summary of statistics for persons and items is presented. The mean and standard deviation values for persons are 0.10 and 0.60, respectively. The person separation value is 1.69, which means that respondents can be grouped into two levels of ability, including the fairly good category ([Aziz et al., 2022](#)). The reliability value for the persons in Table 1 is 0.74, indicating that respondents provided relatively consistent responses to the items in the I-WPECT. This reliability value for persons is considered quite good ([Natsir et al., 2022](#)). Meanwhile, the value of 0.81 is in the good category ([Cerri et al., 2023](#)). Cronbach's Alpha is used to measure overall reliability, namely the extent to which the interaction between respondents and all items in the instrument is consistent. Thus, the interaction between respondents and items in the I-WPECT instrument can be considered reliable.

Item reliability refers to the extent to which an item provides stable and consistent results when used repeatedly on the same subject ([Kurniawan et al., 2024](#); [Zayrin et al., 2025](#)). In this study, the item reliability value was 0.87, indicating that the items in the I-WPECT instrument had high internal consistency. The mean and standard deviation values for each item were 0.00 and 0.23, respectively. The item separation value was 2.64, which means that the items can be grouped into three levels of difficulty. A separation value ≥ 2 indicates that the items in the I-WPECT are quite good at distinguishing the levels of difficulty between questions.

Item difficulty distribution

The difficulty distribution of questions on the I-WPECT instrument is visualized in Figure 1.

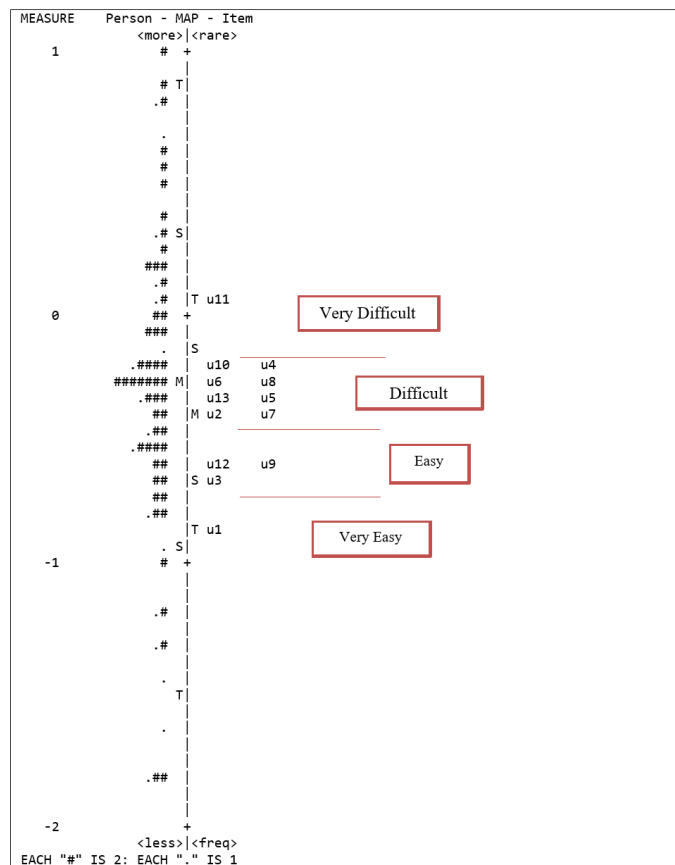


Figure 1 Visualization of the difficulty level of items on the I-WPECT instrument

The distribution of item difficulty levels refers to the distribution of item difficulty levels based on logit values ([Karlimah, 2022](#)). Figure 1 illustrates the relationship between respondent ability (person) and item difficulty level, where the symbol “#” represents two respondents and the symbol “.” represents one respondent. On the right side of the figure (item distribution area), it can be seen that the difficulty level of items is divided into four categories, namely: very difficult, difficult, easy, and very easy. There is one item (7.6%) in the very difficult category, namely item u11. Items in the difficult category comprise 61% of the total, namely items u10, u4, u6, u8, u13, u5, u2, and u7. Furthermore, 23% of items are in the easy category, namely u3, u9, and u12. The other item (7.6%), namely u1, is in the very easy category. Thus, the items in the I-WPECT instrument are distributed across all levels of difficulty, with the majority in the difficult category. An instrument is considered good if its items are evenly distributed across all levels of difficulty ([Reño, 2024](#)), as this allows for a more comprehensive measurement of the various ability levels of respondents.

Item Fit

The item fit of the I-WPECT instrument against the Rasch model is summarized in Table 3.

Table 3 Item fit of the I-WPECT instrument against the Rasch model

No	Item Code	Measure	Model S.E	Infit MNSQ	Outfit MNSQ
1	u1	-0.47	0.08	1.17	1.32
2	u2	0.03	0.08	1.00	1.33
3	u3	-0.28	0.08	1.21	1.27
4	u4	0.18	0.08	0.79	0.73
5	u5	0.06	0.08	1.10	1.13
6	u6	0.12	0.08	0.64	0.64
7	u7	0.01	0.08	0.85	0.85
8	u8	0.10	0.08	0.92	1.10
9	u9	-0.22	0.07	1.07	1.11
10	u10	0.18	0.08	0.69	0.65
11	u11	0.42	0.08	1.19	1.23
12	u12	-0.21	0.07	1.18	1.17
13	u13	0.08	0.08	1.34	1.52

The criteria used to assess the suitability of items with Rasch modeling are determined based on the MNSQ infit and outfit values. According to [Aqida \(2023\)](#), acceptable MNSQ infit and outfit values range from 0.5 to 1.5. In this study, the MNSQ infit values ranged from 0.64 (item u6) to 1.34 (item u13), while the MNSQ outfit values ranged from 0.64 (item u6) to 1.52 (item u13). The MNSQ infit value reflects the suitability of responses to items with a level of difficulty appropriate to the respondent's ability, while the MNSQ outfit value indicates the suitability of responses to items that are considered very easy or very difficult by respondents ([Naingalis et al., 2023](#)). Based on these criteria, there is one item that is categorized as a misfit, namely item u13, because its MNSQ outfit value exceeds the specified upper limit ($1.52 > 1.5$). However, item u13 can still be retained in the instrument because its outfit value does not exceed 2.0, which is still considered within the tolerance limit for instrument development purposes ([Sumintono & Widhiarso, 2015](#)). Thus, 12 of the 13 items in the I-WPECT instrument have good fit to the Rasch model, while 1 item can still be retained.

Unidimensionality

The unidimensionality of the I-WPECT instrument is visualized in Figure 2.

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units				
		Eigenvalue	Observed	Expected
Total raw variance in observations	=	19,6583	100,0%	100,0%
Raw variance explained by measures	=	6,6583	33,9%	34,7%
Raw variance explained by persons	=	1,5646	8,0%	8,2%
Raw Variance explained by items	=	5,0937	25,9%	26,5%
Raw unexplained variance (total)	=	13,0000	66,1%	100,0%
Unexplned variance in 1st contrast	=	2,4028	12,2%	18,5%
Unexplned variance in 2nd contrast	=	2,1218	10,8%	16,3%
Unexplned variance in 3rd contrast	=	1,3690	7,0%	10,5%
Unexplned variance in 4th contrast	=	1,2199	6,2%	9,4%
Unexplned variance in 5th contrast	=	0,9253	4,7%	7,1%

Figure 2 Unidimensionality in the I-WPECT instrument

Unidimensionality is a key assumption in Rasch modeling, which indicates that all items in a scale consistently measure a single primary construct ([Alnahdi et al., 2025](#)). This assumption is considered fulfilled if the raw variance explained by measures is at least 30%, and the unexplained variance in the first contrast does not exceed 15% ([Sumintono & Widhiarso, 2015](#)). Based on the analysis results, the raw variance explained by measures was 33.9%, indicating that about one-third of the total data variance could be explained by the Rasch model. Meanwhile, the unexplained variance in the first contrast is 12.2%, which is still within the tolerance limit. Thus, all items in the I-WPECT instrument reflect a single main dimension: the concept of effort, power, and energy. The I-WPECT instrument adequately meets the assumption of unidimensionality, and its items generally measure the same construct.

Gender DIF

Differential item functioning (DIF) based on gender in the I-WPECT instrument is summarized in Table 4.

Table 4 DIF in the I-WPECT instrument

No	Item Code	DIF Contrast	Rasch-welch t	Prob.
1	u1	-0.43	-2.28	0.0280
2	u2	0.19	1.03	0.3093
3	u3	0.49	2.46	0.0189
4	u4	0.17	0.88	0.3868
5	u5	-0.07	-0.36	0.7181
6	u6	-0.07	-0.36	0.7182
7	u7	0.34	1.81	0.0781
8	u8	0.18	0.93	0.3569
9	u9	0.07	0.38	0.7094
10	u10	-0.21	-1.04	0.3041
11	u11	-0.30	-1.29	0.2073
12	u12	-0.47	-2.41	0.0215
13	u13	0.03	0.16	0.8756

Differential Item Functioning (DIF) analysis aims to identify whether an item in an instrument functions differently when responded to by different groups of respondents, even though they have the same level of ability ([Sumintono & Widhiarso, 2015](#)). In this context, DIF is used to examine potential item bias based on gender differences. If there are differences in response patterns between men and women with equivalent abilities, then the item can be categorized as biased.

Based on the DIF analysis results shown in Table 3, the DIF contrast values range from -0.47 to 0.49. The Rasch-Welch t values range from -2.41 to 2.46, while the probability

values (p-values) range from 0.0189 to 0.8756. Referring to the Rasch-Welch t and p-value, there are three items (u1, u3, and u12) that potentially contain bias, as the t-value is outside the range of -2 to 2 and the p-value is less than 0.05. However, based on the DIF contrast values, these three items are still within the range of -0.5 to 0.5, which is considered neutral (Hayat et al., 2023), so they are not categorized as biased items. Considering these three indicators, it can be concluded that no items in the I-WPECT instrument significantly contain gender bias. This supports that the I-WPECT instrument has met the principle of fairness in measurement in the context of gender.

CONCLUSION

This study aims to adapt and validate a four-tier diagnostic test instrument that measures misconceptions about the concepts of work, power, and energy across cultures using the Rasch modeling approach. Based on the analysis results, the item reliability value was 0.87 and the Cronbach's Alpha value was 0.81, indicating that the instrument has high internal consistency. All items are spread across four levels of difficulty, ranging from very easy to very difficult. Based on the MNSQ infit and outfit values, 12 of the 13 items show good fit with the Rasch model, while the other item can still be retained because its MNSQ outfit value is below the tolerance limit of 2.0. In terms of unidimensionality, all items in the I-WPECT instrument represent one main construct, namely the concept of work, power, and energy. In addition, no potential item bias based on gender was found, based on DIF contrast analysis, Rasch-Welch t, and p-value. Thus, it can be concluded that this four-tier diagnostic test instrument on the topic of work and energy is suitable for identifying misconceptions among high school students in Indonesia.

The limitation of this study is that the instrument testing is still limited to students from public schools, so it does not yet include students from private schools or vocational schools. Therefore, further research is recommended to expand the context of testing this instrument to a more diverse population, including private high schools and vocational schools, to ensure more generalizable results. This research has implications for the availability of diagnostic tests for misconceptions in work, power, and energy materials, which can help educators in Indonesia diagnose students' misunderstandings more accurately and design clearer, more targeted learning interventions.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization, GF and MIS; methodology, GF and MIS; formal analysis, GF and MIS; writing—preparation of the original draft, GF, MIS, TKI; Authorship should be limited to those who have contributed substantially to the reported work.

ACKNOWLEDGMENTS

The authors would like to thank the students and headmaster of SMA Negeri 1 Jetis, Yogyakarta, for allowing them to conduct this research at their school.

REFERENCES

- Alnahdi, A. H., Alsubiheen, A. M., & Aldaihan, M. M. (2025). Rasch measurement model supports the unidimensionality and internal structure of the arabic oswestry disability index. *Journal of Clinical Medicine*, 14(4), 1259. <https://doi.org/10.3390/jcm14041259>
- Anggrayni, S., & Ermawati, F. U. (2019). The validity of Four-Tier's misconception diagnostic test for Work and Energy concepts. *Journal of Physics: Conference Series*,

- 1171, 012037. <https://doi.org/10.1088/1742-6596/1171/1/012037>
- Aqida, D. S., Destia Khairunisa, Meisya, R., & Jannah, R. (2023). The use of rasch model in analyzing grade v elementary school mathematics items. *Sunan Kalijaga International Journal on Islamic Educational Research*, 7(1), 57–67. <https://doi.org/10.14421/skijier.2023.71.06>
- Aripiani, S. K., Samsudin, A., Kaniawati, I., Novia, H., Aminudin, A. H., Sutrisno, A. D., & Coştu, B. (2023). Diagnostic instruments of four-tier test work and energy (fortune) to identify the level of students' conceptions. *Tadris: Jurnal Keguruan Dan Ilmu Tarbiyah*, 8(1), 19–32. <https://doi.org/10.24042/tadris.v8i1.13524>
- Aziz, K. A., Jabar, M. A., Abdullah, S., & Nor, R. N. H. (2022). Assessing factors influencing internet banking adoption by using rasch model measurement. *Bulletin of Electrical Engineering and Informatics*, 11(4), 2313–2321. <https://doi.org/10.11591/eei.v11i4.3761>
- Bawamenewi, A. (2024). Analisis miskonsepsi siswa kelas x sma di kota padang tentang materi hukum newton dengan menggunakan tes diagnostik four-tier multiple choice. *Physics Learning and Education*, 2(2), 89–96. <https://doi.org/10.24036/ple.v2i2.120>
- Bentaleb, K. A., Dachraoui, S., Hassouni, T., Alibrahmi, E. mehdi, Chakir, E., & Belboukhari, A. (2022). Development of a survey to assess conceptual understanding of quantum mechanics among moroccan undergraduates. *European Journal of Educational Research*, volume-11-(volume-11-issue-4-october-2022), 2219–2243. <https://doi.org/10.12973/eu-jer.11.4.2219>
- Cerri, L. Q., Justo, M. C., Clemente, V., Gomes, A. A., Pereira, A. S., & Marques, D. R. (2023). Insomnia Severity Index: A reliability generalisation meta-analysis. *Journal of Sleep Research*, 32(4), 1–12. <https://doi.org/10.1111/jsr.13835>
- Fenditasari, K., Jumadi, Istiyono, E., & Hendra. (2020). Identification of misconceptions on heat and temperature among physics education students using four-tier diagnostic test. *Journal of Physics: Conference Series*, 1470(1), 012055. <https://doi.org/10.1088/1742-6596/1470/1/012055>
- Fuchs, T. T., Bonney, K. M., & Arsenault, M. (2021). Leveraging student misconceptions to improve teaching of biochemistry & cell biology. *The American Biology Teacher*, 83(1), 5–11. <https://doi.org/10.1525/abt.2021.83.1.5>
- Härmälä-Braskén, A.-S., Hemmi, K., & Kurtén, B. (2020). Misconceptions in chemistry among Finnish prospective primary school teachers – a long-term study. *International Journal of Science Education*, 42(9), 1447–1464. <https://doi.org/10.1080/09500693.2020.1765046>
- Hayat, B., Putra, M. D. K., Rahmawati, R., & Isa, K. (2023). Rasch calibration and differential item functioning (dif) analysis of the indonesian national assessment program-language (INAP-L). *World Journal of English Language*, 13(6), 402. <https://doi.org/10.5430/wjel.v13n6p402>
- Irmak, M., İnaltun, H., Ercan-Dursun, J., Yanış-Kelleci, H., & Yürük, N. (2023). Development and application of a three-tier diagnostic test to assess pre-service science teachers' understanding on work-power and energy concepts. *International Journal of Science and Mathematics Education*, 21(1), 159–185. <https://doi.org/10.1007/s10763-021-10242-6>
- Karlimah, K. (2022). How does Rasch modeling reveal difficulty and suitability level the fraction test question? *Jurnal Elemen*, 8(1), 66–76. <https://doi.org/10.29408/jel.v8i1.4170>
- Kurniawan, A., Istiyono, E., & Naba, S. D. (2024). Item quality analysis of physics concept understanding test with rasch model. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, 9(3), 474. <https://doi.org/10.26737/jipf.v9i3.5692>

- Linacre, J. M. (1994). *Sample size and item calibration stability*. Rasch Measurement Transactions. <https://www.rasch.org/rmt/rmt74m.htm>
- Linacre, J. M. (2021). *Winsteps® (Version 4.6.1) [Computer Software]*. <http://www.winsteps.com>
- lv, Q., Zhang, X., Wang, Y., Xu, X., & Zang, X. (2024). Cross-cultural adaptation and validation of the caregiver self-efficacy in contributing to patient self-care scale in China. *BMC Public Health*, 24(1), 1977. <https://doi.org/10.1186/s12889-024-19534-2>
- Maison, Asrial, Susanti, N., Effrita, A., & Tanti. (2021). Identification of students' misconception about light using a four-tier instrument. *Journal of Physics: Conference Series*, 1876(1), 012063. <https://doi.org/10.1088/1742-6596/1876/1/012063>
- Mbonyirivuze, A., Yadav, L. L., & Amadalo, M. M. (2019). Students' conceptual understanding of electricity and magnetism and its implications: A review. *African Journal of Educational Studies in Mathematics and Sciences*, 15(2), 55–67. <https://doi.org/10.4314/ajesms.v15i2.5>
- Naingalis, M. L. P., Fauzi, A., Supurwoko, & Sartono, B. (2023). *Item fit investigation of computational thinking scales instrument among Indonesian physics undergraduate students: A Rasch model analysis*. 090006. <https://doi.org/10.1063/5.0105746>
- Nasir, M., Cari, C., Sunarno, W., & Rahmawati, F. (2023). The development of a two-tier diagnostic test for student understanding of light. *Journal of Turkish Science Education*, 20(1), 150–172. <https://doi.org/10.36681/tused.2023.009>
- Natsir, I., Mayasari, D., & Munfarikhatin, A. (2022). Rasch model: Quality of test instruments to measure students' ability in algebra. *SHS Web of Conferences*, 149, 01033. <https://doi.org/10.1051/shsconf/202214901033>
- Ravish, I. (2025). An analysis of misconceptions in science education: identifying barriers to effective learning and instruction. *International Journal of Applied and Behavioral Sciences*, 2(1), 199–207. <https://doi.org/10.70388/ijabs250119>
- Reño, A. B. (2024). Rasch analysis in diagnosing learning gaps and gauge item difficulty in eighth-grade mathematics assessments. *Pakistan Journal of Life and Social Sciences (PJLSS)*, 22(2), 14571–14579. <https://doi.org/10.57239/PJLSS-2024-22.2.001049>
- Sari, F. K., Alifiani, A., Nursit, I., Jannah, U. R., & Rahaju, R. (2024). Unveiling hidden misconceptions: A new perspective on first-year students' algebraic understanding. *JME (Journal of Mathematics Education)*, 9(2), 226–234. <https://doi.org/10.31327/jme.v9i2.2231>
- Sukarelawan, M. I., Indratno, T. K., Oktova, R., & Abdullah, N. S. Y. (2024). Quality of four-tier diagnostic test on wave and vibration materials: An empirical study using rasch modeling. *Jurnal Pendidikan Fisika Indonesia*, 20(1), 27–37. <https://doi.org/10.15294/jpfi.v20i1.47493>
- Sukarelawan, M. I., Jumadi, J., & Rahman, N. A. (2019). An analysis of graduate students' conceptual understanding in heat and temperature (h&t) using three-tier diagnostic test. *Indonesian Review of Physics*, 2(1), 9. <https://doi.org/10.12928/irip.v2i1.910>
- Sukarelawan, M. I., Sa'adah, M., Jupri, J., & Hikmah, F. N. (2024). Is it true that my students don't understand the static fluid concepts? Rasch modeling perspective. *Jurnal Pendidikan dan Pengajaran*, 57(1), 127-138.
- Sukarelawan, M. I., Sriyanto, S., Puspitasari, A. D., Sulisworo, D., & Hikmah, U. N. (2021). Four-tier heat and temperature diagnostic test (4t-htdt) to identify student misconceptions. *JIPFRI (Jurnal Inovasi Pendidikan Fisika Dan Riset Ilmiah)*, 5(1), 1–8. <https://doi.org/10.30599/jipfri.v5i1.856>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada asesmen pendidikan*. Trim Komunika.
- Suprpto, N. (2020). Do we experience misconceptions?: An ontological review of

- misconceptions in science. *Studies in Philosophy of Science and Education*, 1(2), 50–55. <https://doi.org/10.46627/sipose.v1i2.24>
- Syuhendri, S. (2017). A learning process based on conceptual change approach to foster conceptual change in newtonian mechanics. *Journal of Baltic Science Education*, 16(2), 228–240. <https://doi.org/10.33225/jbse/17.16.228>
- Syuhendri, S. (2019). Student teachers' misconceptions about gravity. *Journal of Physics: Conference Series*, 1185, 012047. <https://doi.org/10.1088/1742-6596/1185/1/012047>
- Syuhendri, S. (2022). Teaching for conceptual change on Newton's First Law. *Journal of Physics: Conference Series*, 2165(1), 012036. <https://doi.org/10.1088/1742-6596/2165/1/012036>
- Zayrin, A. A., Nupus, H., Maizia, K. K., Marsela, S., Hidayatullah, R., & Harmonedi, H. (2025). Analisis instrumen penelitian pendidikan (uji validitas dan reliabilitas instrumen penelitian). *Jurnal QOSIM Jurnal Pendidikan Sosial & Humaniora*, 3(2), 780–789. <https://doi.org/10.61104/jq.v3i2.1070>

AUTHOR PROFILES



Gress Faihah is a student in the Department of Physics Education, Faculty of Teacher Training and Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia. She can be contacted at gress2100007013@webmail.uad.ac.id



Dr. Moh. Irma Sukarelawan, M.Pd is an assistant professor at Universitas Ahmad Dahlan (Department of Physics Education, Faculty of Teacher Training and Education), Yogyakarta, Indonesia. His research focuses on physics education, misconception, metacognition, and Rasch modelling. He can be contacted at irma.sukarelawan@pfis.uad.ac.id [Google Scholar](#), [Sinta ID](#), [Orcid](#), [Scopus](#).



Toni Kus Indratno, M.Pd.Si is an associate professor at Universitas Ahmad Dahlan (Department of Physics Education, Faculty of Teacher Training and Education), Yogyakarta, Indonesia. His research focuses on physics education and physics learning technology. He can be contacted at tonikus@staff.uad.ac.id [Google Scholar](#), [Sinta ID](#), [Orcid](#), [Scopus](#).