



DEVELOPMENT AND VALIDATION OF A TWO-TIER FORMAL REASONING TEST INSTRUMENT FOR HIGH SCHOOL STUDENTS ON THE TOPIC OF ROTATIONAL MOTION

Windy Nerissa¹, Ridwan Efendi^{2*}, Dedi Sasmita³

^{1,2,3} Program in Physical Education, FPMIPA, University of Education Indonesia
Jl. Dr. Setiabudhi No. 229, Bandung
ridwanefendi@upi.edu^{1*}

Received: 24 06 2026. Accepted: 20 07 2026. Published: 07 2026

Abstrak

Pembelajaran fisika keterampilan berpikir saintifik yang mendalam, terutama pada materi kompleks seperti dinamika rotasi. Sayangnya, instrumen evaluasi yang ada masih didominasi kalkulasi matematis tinggi dan format satu tingkat yang rentan tebakan acak. Penelitian ini bertujuan untuk mengembangkan dan memvalidasi instrumen tes diagnostik penalaran formal kualitatif berformat Two-Tier Multiple Choice yang terbebas dari intervensi komputasi matematis. Penelitian eksploratif ini menggunakan pendekatan Research and Development (R&D) yang disesuaikan dari tahapan rencana tes Djemari Mardapi. Instrumen yang dikembangkan terdiri dari 11 pasang pilihan ganda dua tingkat dengan total 22 butir soal. Uji validitas melibatkan 7 orang panel ahli, sedangkan uji coba empiris dilakukan pada 75 siswa menengah atas menggunakan teknik purposive sampling. Hasil penelitian menunjukkan: (1) validitas isi memiliki rentang indeks Aiken's V 0,84 – 1,00 (kriteria sangat baik); (2) validitas empiris membuktikan seluruh butir soal valid dengan rentang korelasi Pearson 0,324 – 0,598; (3) reliabilitas instrumen berada pada kategori reliabel (Cronbach's Alpha = 0,647); (4) tingkat kesukaran didominasi kategori sedang (10 butir) dan sukar (1 butir); serta (5) indeks daya pembeda memadai (0,210 – 0,460), dengan perbaikan redaksional pada dua butir soal (Q2 dan Q5). Kesimpulannya, instrumen tes penalaran formal dinamika rotasi terbukti valid dan reliabel untuk mengukur kapasitas kognitif siswa tanpa hambatan operasional matematis.

Kata Kunci : Dinamika rotasi; penalaran formal; pilihan ganda dua tingkat; itu; validitas

Abstract

Physics learning requires deep scientific thinking skills, especially in complex topics like rotational dynamics. Unfortunately, current evaluation instruments are still dominated by single-tier formats prone to random guessing and mathematical calculation bias. This study aims to develop and validate a Two-Tier Multiple Choice diagnostic instrument free from mathematical interventions and analyzed using a polytomous Rasch model (Partial Credit Model / PCM). This exploratory research employs a Research and Development (R&D) approach based on Djemari Mardapi's model. The developed instrument consists of 11 paired questions (22 items). Content validity involved 7 expert panels, while empirical testing was conducted on 75 high school students (N=75). Data were analyzed using a 4-category polytomous model via Winsteps to evaluate empirical validity, dimensionality, and item fit. The results show that the instrument meets the essential unidimensionality requirement (1st contrast eigenvalue = 1.68). The calibration yielded ideal average infit and outfit MNSQ values (1.00 and 1.04) and a strong item reliability of 0.80. The monotonic progression of the average logit across the four scoring categories (-0.18, -0.06, 0.12, 0.43 logits) empirically validated the efficacy of the scoring structure. In conclusion, the application of polytomous Rasch analysis to two-tier tests accurately diagnoses students' formal reasoning barriers without being distorted by basic mathematical flaws.

Keywords: Rotational dynamics; formal reasoning; two-tier multiple choice; partial credit model; rasch modeling

INTRODUCTION

In high schools, physics education should not merely focus on transferring algorithmic knowledge. Instead, it should help students develop deep scientific thinking skills (Zimmerman, 2000).

One of the areas of mechanics that requires a high level of conceptual understanding is the dynamics of rotation and the equilibrium of rigid bodies. This topic is particularly challenging because it involves integrating various physical concepts simultaneously, including moments of inertia, torque, and the law of conservation of angular momentum.

Consistent with the findings Isnaini et al. (2025), difficulties in understanding these concepts are often due to a lack of deep conceptual understanding and the use of inappropriate problem-solving strategies. The main challenge for students isn't necessarily an inability to perform mathematical operations, but rather a weakness in conceptualizing the abstract relationships between variables without relying on formulas (Verlinda & Latifah, 2020). When faced with qualitative analyses related to the motion of objects or changes in mass distribution, students tend to rely on incorrect initial assumptions or even develop persistent misconceptions (Rahmawati & Zulaikah, 2017).

Students' limited ability to solve conceptual problems is closely related to their stage of cognitive development. According to theories of intellectual development, high school students should ideally have reached the formal operational stage of thinking. This stage involves the ability to think hypothetically and deductively, as well as to make proportional judgments (Inhelder & Piaget, 1958; Zimmerman, 2000). To effectively assess whether students have

reached this stage of cognitive development, specific assessment tools are necessary, such as the Lawson Classroom Test of Scientific Reasoning (LCTSR) (Lawson, 1978).

However, large-scale psychometric evaluations confirm that there are inconsistencies in the cognitive structures of learners (Bao et al., 2018). In fact, this phenomenon of delayed cognitive development persists even at the university level (Hrouzková, T., & Richterek, 2021). In Indonesia, most students remain at the concrete and transitional stages of reasoning (Ariadi et al., 2025; Mayasyafira et al., 2025). The need to improve the quality of reasoning is further compounded by the limitations of the evaluation tools used in schools. After all, the accuracy of diagnostic tools is crucial for objectively assessing learners' cognitive structures without bias (Eryilmaz et al., 2015). Single-choice tests are particularly susceptible to guessing errors (Tukiyo et al., 2023). Even worse, although the Two-Tier test was developed to address these issues (Lengkong et al., 2021), it still has its own limitations. In 2021, many local studies still analyzed two-tiered data using traditional dichotomous approaches. These approaches failed to accurately capture the nuances in reasoning patterns. In reality, such measurements require a polytomous approach (Kassiavera et al., 2024).

The focus of this research is to develop formal reasoning instruments in the form of two-tier multiple-choice questions, with these instruments being applied to the context of rotational motion without the need for mathematical formulas. This research introduces a novel methodological approach by utilizing Rasch model theory. This theory allows for a detailed calibration of the reasoning process, by evaluating both the

accuracy of students' multiple-choice answers and the validity of their reasoning. This approach enables the systematic identification of cognitive barriers (Kassiavera et al., 2024).

This instrument utilizes various real-world phenomena (such as yo-yos, door hinges, gears, helicopter rotors, and carnival ride components) to avoid the biases associated with theoretical memorization from textbooks. Specifically, this study aims to comprehensively evaluate the psychometric properties of this instrument by conducting content validity tests and examining the empirical characteristics of its items using modern methods. Additionally, the study seeks to explore students' ability to think logically using this diagnostic tool, by analyzing the data using the Rasch partial credit model.

The specific objectives of this research include: (1) evaluating the content validity and empirical reliability of the instrument; (2) calibrating the difficulty level of questions and ensuring their appropriateness for the purpose of assessment; and (3) validating the scoring structure used to categorize students' reasoning abilities. The research questions are: (1) What are the reliability and validity levels of the two-tier dynamic rotation instrument?; (2) To what extent can the Rasch model effectively capture the functioning of the scoring categories and question appropriateness?

METHODS

This study employs a quantitative approach, with an exploratory and developmental research design. The development procedure was adapted from the stages outlined in Mardapi's Model (Mardapi, 2012). This method involves five systematic steps to ensure reproducibility of

the research: (1) determining the test specifications by breaking down key indicators into operational indicators based on contextual factors; (2) drafting 11 pairs of questions each, for a total of 22 questions; (3) validating the content of the questions through expert judgment; (4) conducting a pilot study; and (5) analyzing the questions quantitatively.



Figure 1. Process of developing test instruments

The target population in this study consists of students in the 11th grade of science and mathematics programs in high schools, who have already studied the concepts related to rotational dynamics. For the pilot test, a combination of purposive sampling and quota sampling was used. As a result, 75 respondents from various high schools were selected. This sample size was determined based on the fact that this study is still in the pilot phase. Methodologically, 75 respondents is a sufficient number to evaluate the basic parameters of the test questions.

The scope of the instrument was developed in a *Two-Tier Multiple Choice format* totaling 11 pairs with a total of 22 questions that measured formal reasoning indicators, namely Proportional Reasoning, Control of Variables, Probabilistic Reasoning, Correlational Reasoning, and Hypothetico-Deductive Reasoning.

The data collection tools included Likert-scale questionnaires rated on a 1-4 scale, designed to assess various aspects: item content/relevance and representativeness, measurement model/construct alignment,

item style/clarity, and comprehensiveness of the items (Grant & Davis, 1997).

Data analysis was conducted in two stages. First, the data provided by 7 expert validators (4 physics education professors, 1 physics professor, and 2 physics teachers) were analyzed using Aiken's V index (Aiken, 1985) to assess the validity of the content. Second, empirical data analysis was carried out using the Winsteps program, based on the Rasch model with the Partial Credit Model (PCM) approach. Response patterns were categorized into 4 hierarchical categories (Kassiavera et al., 2024).

Table 1. Patterns of Two-Tier Answer Combinations Among Students

Answer Category	Combination (Tier 1 - Tier 2)	Score
Scientific Reasoning (SR)	Correct - Correct	4
Guessing/Error	Incorrect - Correct	3
Error/Misconception	Correct - Incorrect	2
No Reasoning	Incorrect - Incorrect	1

Statistical analysis of the items in this questionnaire evaluates their reliability, difficulty level, and item fit (*item fit* / *infit* & *outfit* MNSQ) This helps ensure the empirical validity of the instrument, free from any mathematical calculation biases (Duman & Dökme, 2024; Kassiavera et al., 2024).

RESULTS AND DISCUSSIONS

The quality of this instrument for measuring formal reasoning was assessed through content validity tests and empirical field trials, in order to obtain quantitative data regarding the characteristics of each test item. Based on the analysis of data from 75 respondents using the Rasch partial credit

model, the following psychometric findings were obtained:

1. Content Validity Score (Aiken's V)

The validity of the content is intended to assess the theoretical soundness of the instrument. Based on the evaluations of 7 experts (professors and teachers) regarding the preliminary version of the instrument, which consisted of 11 pairs of items, totaling 22 items overall, the validity was measured using Aiken's V formula. This evaluation covered various aspects: item content/relevance and representativeness, measurement model/construct alignment, item style/clarity, and comprehensiveness of the items (Grant & Davis, 1997). The results of the Aiken's V calculations are presented in Table 1.

Table 2. Summary of Aiken's V Content Validity Results

Question Item No.	Σs	Aiken's Value (V)	Criteria
Q1	19.7	0.94	Valid
Q2	17.7	0.84	Valid
Q3	21.0	1,00	Valid
Q4	19.0	0.90	Valid
Q5	18.7	0.89	Valid
Q6	19.0	0.90	Valid
Q7	18.3	0.87	Valid
Q8	18.7	0.89	Valid
Q9	19.3	0.92	Valid
Q10	19.7	0.94	Valid
Q11	18.3	0.87	Valid

According to Table 1, the Aiken's V index values for all 11 items range from 0.84 to 1.00. All items consistently met the minimum threshold required by the raters' agreement criterion ($V \geq 0.75$).

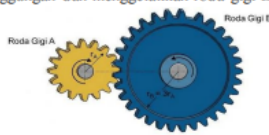
To visualize the characteristics of instruments that focus on logical reasoning and are free from the burden of complex mathematical calculations, two representations of the questions in the Two-

Tier Multiple Choice format are presented in Figures 2 and 3.

1. Roda gigi A (benukuran kecil) saling bersinggungan dan menggerakkan roda gigi B (benukuran besar). Jari-jari roda gigi B adalah tiga kali lipat lebih besar daripada jari-jari roda gigi A. Jika roda gigi A diputar sebanyak 12 putaran penuh, berapa kalikah roda gigi B akan berputar?
- A. 36 kali
B. 12 kali
C. 4 kali
D. 9 kali
E. 15 kali
2. Alasan:
- A. Roda gigi yang lebih besar memiliki keliling lintasan yang lebih panjang, sehingga dalam satu putaran ia akan menghasilkan jumlah putaran yang lebih banyak pada roda lawannya..
B. Jumlah putaran roda gigi berbanding lurus dengan ukuran jari-jarinya.
C. Jari-jari roda B tiga kali lipat lebih besar, maka dibutuhkan tiga putaran roda A untuk menghasilkan satu putaran roda B.
D. Ukuran roda gigi tidak memengaruhi jumlah putaran, melainkan hanya mengubah arah putaran.
E. Kedua roda saling mengunci satu sama lain, sehingga jumlah putarannya pasti sama persis.

Figure 2. Design of the two-tier question format for the indicator of proportional reasoning (Gear System)

As shown in Figure 2, the instrument is designed to utilize visual stimuli that challenge students' understanding of physical ratios and proportions. In question Q1, students are presented with two gears that are in contact with each other. The first-level questions require students to predict the number of rotations of gear B based on the rotations of gear A. The second-level questions assess the logical reasoning behind students' answers. Students must understand that since the radius of gear A is three times larger, gear A must rotate three times as many times to cause gear B to complete one full rotation.



17. Seorang siswa mengamati dua jenis Yoyo dengan massa total yang sama persis: Yoyo 'Kupu-kupu' yang beban massanya ditempatkan di velg terluar (pinggir), dan Yoyo Klasik yang massanya terpusat di tengah (poros). Saat dimainkan dengan gaya tarikan yang sama, siswa tersebut melihat bahwa Yoyo Kupu-kupu mampu berputar di ujung tali jauh lebih lama dibandingkan Yoyo Klasik. Ia ingin meneliti secara ilmiah fenomena letak massa tersebut. Berdasarkan observasi di atas, manakah rumusan hipotesis yang paling tepat secara prinsip dinamika rotasi?



- A. Jika yoyo kupu-kupu dan yoyo klasik dimainkan secara bersamaan, maka yoyo kupu-kupu pasti akan berputar lebih lama di ujung tali.
B. Semakin jauh beban massa didistribusikan dari poros putar, semakin besar hambatan putarnya (momen inersia), sehingga yoyo mampu mempertahankan putarannya lebih lama.
C. Desain yoyo kupu-kupu yang lebar bertujuan mengurangi gesekan dengan udara, sehingga putarannya menjadi jauh lebih awet daripada yoyo klasik.
D. Massa total yang dimiliki sebuah yoyo berbanding lurus dengan durasi putaran yoyo tersebut saat berada di ujung tali.
E. Jika beban massa terpusat di poros, maka gaya gravitasi bumi akan menarik yoyo dengan lebih kuat sehingga yoyo cepat berhenti berputar.
18. Alasan:
- A. Karena berdasarkan fenomena pengamatan tersebut, peletakkan distribusi massa mempengaruhi lamanya waktu putaran yang terjadi, sehingga pernyataan ini menjadi penjelasan mekanis yang dapat diuji kebenarannya.
B. Karena hipotesis ilmiah cukup menyusun struktur tebakan langsung mengenai yoyo mana yang akan menang (berputar lebih lama), tanpa harus memberikan penjelasan mekanis di baliknya.
C. Karena berdasarkan pengamatan visual, bentuk yoyo kupu-kupu yang lebar diyakini meminimalkan gesekan udara, sehingga udara menjadi satu-satunya variabel yang menentukan awetnya putaran..
D. Karena memindahkan letak massa ke bagian pinggir sering disalahpahami memberikan efek beban total yang lebih berat, sehingga tarikan gravitasi menjadi penentu durasi putarannya.
E. Karena pernyataan tersebut sudah merupakan hukum mutlak dalam dinamika rotasi, sehingga fenomena tersebut hanyalah pembuktian teori yang tidak perlu lagi diuji melalui eksperimen.

Figure 3. Design of the two-tier question format for the Hypothetical-Deductive Reasoning indicator (Yoyo phenomenon)

In addition to proportional reasoning, this tool also assesses students' ability to analyze more complex phenomena related to rotational dynamics, as shown in Figure 3.

According to Figure 3, the instrument is designed to use visual stimuli that challenge students' understanding of real physical phenomena. In question Q9, students are asked to observe two types of yo-yos with exactly the same total mass. In the "Butterfly" yo-yo, the mass is distributed around the outer rim of the yo-yo, while in the traditional yo-yo, the mass is concentrated at the center

of the axis. By explaining why the Butterfly yo-yo can spin longer on the end of a string when subjected to the same pulling force, students are required to determine the most appropriate hypothesis based on principles of rotational dynamics.

In general, the structure of the questions at the Tier-1 and Tier-2 levels, as shown in Figures 2 and 3, is carefully designed. These questions cannot be answered by simply guessing or memorizing mathematical formulas like $I = mr^2$. Instead, these questions require students to utilize their formal operational reasoning skills. The goal is to enable students to analyze and relate various physical variables within the context of real-world problems.

Conceptually, the high validity of these results indicates that the instrument's design, which eliminates the need for mathematical calculations, has successfully captured the essential aspects of rotational dynamics. The decision to replace traditional laboratory objects with real-world phenomena like yo-yos, gears, and door hinges proved effective in minimizing memory-based biases. These findings are consistent with the principle of reducing cognitive load, which states that simplifying arithmetic tasks helps students and evaluators focus entirely on assessing the logical structure of hypotheses (Açikgöl, 2021). The instrument's design also meets the need to assess higher-level abstract thinking skills, without being affected by students' basic mathematical errors (Verlinda & Latifah, 2020).

2. Overall Calibration and Instrument Reliability

The results of calibrating the two-tier diagnostic instrument are presented in Table 1. These statistics provide an overview of the reliability and goodness-of-fit of the model.

Table 3. Summary Statistics from The Polytomous Rasch Model

Statistic	Person (N=75)	Item (N=11)
Measure (Logit)		
Mean	0.13	0.00
SD	0.49	0.24
Infit MNSQ		
Mean	1.00	1.00
SD	0.25	0.17
Outfit MNSQ		
Mean	1.04	1.04
SD	0.41	0.27
Separation & Reliability		
Separation Index	1.21	2.03
Reliability	0.60	0.80
Cronbach's Alpha	0.61	-

According to Table 3, the instruments demonstrate extremely precise measurement results. The average values of Infit MNSQ and Outfit MNSQ for both individuals and items are very close to the ideal value of 1.00. This indicates that all test items fit well with the Rasch model, with no significant deviations or anomalies. These findings are consistent with literature that suggests that MNSQ values around 1.00 indicate that data can be accurately predicted by the measurement model (Cvenic et al., 2022; Kassiavera et al., 2024)

The separation reliability for the individual test items was 0.80 (with a separation index of 2.03), indicating that the instrument has a strong ability to distinguish between different levels of difficulty among the test items. The reliability of the participants was 0.60, and Cronbach's Alpha was 0.61. These values are considered satisfactory for a pilot study that aims to measure complex and transitional aspects of reasoning.

3. Dimensionality of the Instrument

Principal Component Analysis (PCA) was performed on the residuals to confirm the unidimensionality of the instruments. The results are presented in Table 4.

Table 4. Standardized Residual Variance (PCA)

Variance	Eigenvalue	Observed (%)
Raw variance explained by measures	3.70	25.2%
Raw variance explained by persons	2.78	18.9%
Raw variance explained by items	0.93	6.3%
Raw unexplained variance (total)	11.00	74.8%
Unexplained variance in 1st contrast	1.68	11.4%

According to Table 4, the variance that can be explained by the various measures amounts to 25.2%. More importantly, the residual variance, which cannot be explained by these measures, is very low at 11.4%, with an eigenvalue of 1.68. Since the eigenvalue is below the critical value of 2.0, this instrument meets the requirements for unidimensionality, as confirmed empirically by Kassiavera et al. (2024). This finding is crucial and consistent with the recommendations of Bao et al. (2018), who emphasized that scientific reasoning tests must be unidimensional in nature. In this case, the low eigenvalue indicates that the instrument truly measures only one dimension: formal reasoning skills related to rotational dynamics, without any interference from other dimensions such as students' mathematical computation abilities (Isnaini et al., 2025).

4. Validity of the Polytomous Scoring Structure (Partial Credit Model)

The application of the 4 categories of the PCM (Partial Credit Model) demonstrates that the instrument is effective at capturing students' cognitive transitions (Inhelder & Piaget, 1958). A summary of the category structure is presented in Table 5.

Table 5. Summary of Category Structure (Partial Credit Model)

Score	Category of Answers	Observed Count	Observed (%)	Observed Average (Logit)	Infit MNSQ	Outfit MNSQ
1	No Reasoning	267	32%	-0.18	1.03	1.00
2	Misconception	97	12%	-0.06	0.92	1.54
3	Guessing	120	15%	0.12	1.13	1.32
4	SR	341	41%	0.43	0.95	0.94

The frequency distribution shown in Table 5 indicates that Category 4 was the most common response, accounting for 41%. Category 1 came in second, with 32%. Responses that represented partial guesses Categories 2 and 3 accounted for 12% and 15%, respectively.

The analysis of the observed average scores shows a systematic increase across categories: -0.18 logits (Category 1), -0.06 logits (Category 2), 0.12 logits (Category 3), and 0.43 logits (Category 4). This linear increase in logits serves as crucial empirical evidence that the politomical scoring framework effectively reduces the role of guessing in test performance. Students who simply guessed the correct answers without any logical reasoning (Category 2) had significantly lower latent abilities (-0.06 logits) compared to those who made mistakes but used logical reasoning to arrive at their answers (Category 3, 0.12 logits). These findings strongly support the hypotheses proposed by Bao et al. (2018) dan Kassiavera et al. (2024), who argued that two-tier tests combined with politomical analysis can effectively distinguish between genuine misunderstandings and mere guesses, which often go unnoticed in dichotomous tests.

5. Item Fit and Difficulty Evaluation

The mapping of difficulty levels shows that the questions are appropriately structured to distinguish between students at different stages of reasoning, from concrete to formal operational thinking. The details regarding the suitability of each question are presented in Table 6.

Table 6. Item Statistics and Fit (Sorted from Most Difficult)

Item	Measure (Logit)	Infit MNS Q	Outfit MNS Q	Point Measure Correlation (PTMEA)
Q6	0.37	0.87	0.80	0.49
Q5	0.18	1.40	1.56	0.27
Q9	0.18	0.96	0.94	0.43
Q11	0.16	0.98	0.92	0.48
Q3	0.00	0.81	0.76	0.56
Q7	0.00	0.78	0.80	0.47
Q10	-0.01	0.86	0.80	0.60
Q4	-0.02	1.08	1.05	0.47
Q1	-0.09	1.08	1.13	0.34
Q8	-0.20	1.18	1.10	0.41
Q2	-0.59	0.95	1.56	0.29

According to Table 4, Q6 had the highest cognitive load among all questions (0.37 logits), while Q2 was the easiest question to answer (-0.59 logits). The high difficulty level of these questions is reasonable and consistent with the findings of Verlinda & Latifah, (2020). They noted that the concepts related to rotational dynamics require a high level of abstraction, as students must analyze the relationships between multiple variables qualitatively. Without any mathematical formulas to memorize, students often struggle to think logically. This explains why the difficulty level increases when students are forced to rely solely on logical reasoning (Rahmawati & Zulaikah, 2017).

All items also exhibited a positive Point Measure of Agreement (PTMEA), indicating that each item contributed effectively to the overall measurement construct. Additionally, the ranges of Infit and Outfit values for the MNSQ overall indicated a very good fit of the model to the data. This is consistent with Rasch's calibration standards, which require that the instrument function without any anomalies or linguistic deviations that could mislead students (Kassiavera et al., 2024).

CONCLUSION

This research has successfully developed and validated a comprehensive diagnostic tool for assessing qualitative formal reasoning, using a two-tier multiple-choice format related to the concept of rotational dynamics. Through theoretical evaluations and empirical analyses using the Rasch partial credit model with 4 categories, a precise cognitive mapping was obtained, free from distortions caused by mathematical calculations. The tool proved to be valid, unidimensional, and reliable. The increasing logit scores across categories indicate that the Rasch model effectively distinguishes different levels of understanding.

The main scientific contribution of this research lies in the design of a contextualized instrument that eliminates the need for complex mathematical calculations and interventions. This makes it possible to accurately measure the transition from operational to formal logical thinking among students. In practical terms, educators can use this instrument in physics lessons to identify misconceptions and objectively assess students' cognitive barriers. Since this research is still in the pilot study phase, with a limited sample size, it is highly recommended that future studies be conducted on a larger scale. This will help establish standardized benchmarks for scientific reasoning across

the country, as well as explore any correlations with demographic factors such as gender and socioeconomic status.

ACKNOWLEDGMENTS

The author would like to express gratitude to the supervisor for all the guidance, discussions, and support provided throughout the research process. Thanks are also extended to the physics teachers and all the students at the schools where the tests were conducted for their active participation. Additionally, thanks go to the expert reviewers who provided valuable feedback to help improve the test instruments.

REFERENCES/ BIBLIOGRAPHY

- Açikgöl, K. (2021). Developing a Two-Tier Proportional Reasoning Skill Test: Validity and Reliability Studies. *International Journal of Assessment Tools in Education*, 8(2), 357–375. <https://doi.org/10.21449/ijate.909316>
- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Ariadi, P., Dinata, C., Lukas, R., Hartanto, T. J., & Sari, D. K. (2025). *Scientific Reasoning Ability of High School Students in Palangka Raya in Physics Learning*. 12(5), 1438–1452.
- Bao, L., Xiao, Y., Koenig, K., & Han, J. (2018). Validity evaluation of the Lawson classroom test of scientific reasoning. *Physical Review Physics Education Research*, 14(2), 20106. <https://doi.org/10.1103/PhysRevPhysEduRes.14.020106>
- Cvenic, K. M., Planinic, M., Susac, A., Ivanjek, L., Jelacic, K., & Hopf, M. (2022). Development and validation of the Conceptual Survey on Wave Optics. *Physical Review Physics Education Research*, 18(1), 10103. <https://doi.org/10.1103/PhysRevPhysEduRes.18.010103>
- Duman, Z., & Dökme, İ. (2024). SCIENTIFIC REASONING SKILLS TEST DEVELOPMENT STUDY FOR SECONDARY SCHOOL 8TH GRADE STUDENTS. *International New York Academic Research Congress*, 145–152.
- Eryılmaz, A., East, M., & McDermott, L. C. (2015). A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science. 11(5), 989–1008. <https://doi.org/10.12973/eurasia.2015.1369a>
- Grant, J. S., & Davis, L. L. (1997). Selection and Use of Content Experts for Instrument Development. *Research in Nursing & Health*, 20(3), 269–274.
- Hrouzková, T., & Richterek, L. (2021). *Lawson classroom test of scientific reasoning at entrance university level*. 74–85.
- Inhelder, B., & Piaget, J. (1958). The Structural Integration of Formal Thought. In *The Growth of Logical Thinking: From childhood to adolescence* (pp. 245–350).
- Isnaini, F., Tiur, H., Silitonga, M., Musa, M., Hidayatullah, S., Sirait, J., & Afrizon, R. (2025). Diagnosing Students' Problem-Solving Challenges in Rotational Dynamics Using Two-Tier AR Flashcard Tests. *Jurnal Inovasi Pendidikan IPA*, 11(2), 402–417. <https://journal.uny.ac.id/index.php/jipi/article/view/80461>
- Kassiavera, S., Suparmi, A., Cari, C., & Sukarmin, S. (2024). Application of Rasch Model in Two-Tier Test for Assessing Critical Thinking in Physics Education. *Journal of Baltic Science Education*, 23(6), 1227–1242. <https://doi.org/10.33225/jbse/24.23.1227>
- Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1), 11–24.
- Lengkong, M., Istiyono, E., Rampean, B. A. O., Rejeki, A. M., & Nirmala, M. F. T. (2021). *Development of Two-Tier Test Instruments to Detect Student's Physics Misconception*. 528(Icriems 2020), 561–566.

- Mardapi, D. (2012). Pengukuran, penilaian, dan evaluasi pendidikan. *Yogyakarta: Nuha Medika*, 45.
- Mayasyafira, S. D., Ekawati, E. Y., & Astuti, L. D. (2025). Profiles of Science Reasoning from Lawson's Perspective: A Comparative Study of Gender, School Location, and Practicum Experience. *Jurnal Pendidikan Fisika*, 13(3), 363–383. <https://doi.org/10.26618/0vxb4t98>
- Rahmawati, I., & Zulaikah, S. (2017). *Jurnal Pendidikan IPA Indonesia ANALYSIS OF STUDENTS ' DIFFICULTIES ABOUT ROTATIONAL*. 6(1), 95–102. <https://doi.org/10.15294/jpii.v6i1.9514>
- Tukiyo, Efendi, M., Solissa, E. M., Klaten, W. D., Langsa, I., Ambon, U. P., Banyuwangi, N., Sultan, U. I. N., Muhammad, A., & Uinsi, I. (2023). *THE DEVELOPMENT OF A TWO-TIER DIAGNOSTIC TEST TO DETECT STUDENT ' S MISCONCEPTIONS IN LEARNING PROCESS*. 5(1), 1–5.
- Verlinda, S., & Latifah, E. (2020). *The Effectiveness of Modeling Instruction Learning on Students' Conceptual Understanding of Rotational Dynamic*. 10(02), 158–172. <https://doi.org/10.26740/jpfa.v10n2.p158-172>
- Zimmerman, C. (2000). The Development of Scientific Reasoning Skills. *Developmental Review*, 20(1), 99–149. <https://doi.org/https://doi.org/10.1006/dre v.1999.0497>