DOI: 10.14421/impulse.2026.61-03

IMPULSE: Journal of Research and Innovation in Physics Education
Volume 6, Issue 1, 28 – 42

© ISSN (p): 2798–1762; ISSN (e): 2798-1754
http://ejournal.uin-suka.ac.id/tarbiyah/impulse

Rasch Model Analysis of Senior High School Physics Test Items in an Academic Competition in North Sulawesi

Awal Mulia Rejeki Tumanggor*, Parno Sumanro Mahulae, Cristi Ascika Sekeon, Yohansen Frando Hadinata Silaban, Liza Septia Ahmad, Isna Pujianti

Physics Education Study Program, Universitas Negeri Manado, Tondano, Indonesia
*Corresponding author: awaltumanggor@unima.ac.id

ABSTRACT

Evaluating the quality of test items is crucial for accurately assessing student competence in academic competitions such as physics olympiads. This study aims to analyze the quality of physics olympiad test items used in the Temu Fisika 2024 competition for senior high school students across North Sulawesi. A total of 40 multiple-choice items from the preliminary round of the Temu Fisika 2024 physics olympiad were analyzed. The responses of 35 high school students who participated in the competition were examined using the Rasch model through the QUEST program. The analysis covered item fit, reliability (item and person), difficulty level, and distractor effectiveness as complementary CTT analysis. The results showed that 33 items (82.5%) fit the Rasch model, while seven items showed misfit. The Rasch reliability indices were good (Item Reliability = 0.82, Person Reliability = 0.84). The difficulty levels were appropriately distributed, with 50% of items categorized as moderate, 22.5% easy, 20% difficult, 2.5% very easy, and 5% very difficult. However, distractor analysis revealed a significant shortcoming: only 15% of items had all distractors functioning effectively. This study is the first Rasch analysis of the Temu Fisika olympiad, providing empirical evidence that while the test is generally reliable, specific improvements are needed. The findings highlight the importance of revising misfitting items and non-functional distractors to improve the discriminating power of the test, particularly among high-ability participants. This study also provides a practical reference for improving item development and validation processes in academic competitions.

INTISARI

Evaluasi kualitas butir soal sangat penting untuk menilai kompetensi siswa secara akurat dalam kompetisi akademik seperti olimpiade fisika. Penelitian ini bertujuan untuk menganalisis kualitas butir soal olimpiade fisika yang digunakan dalam kompetisi Temu Fisika 2024 untuk SMA di seluruh Sulawesi Utara. Sebanyak 40 butir soal pilihan ganda dari babak penyisihan olimpiade Temu Fisika 2024 dianalisis. Jawaban dari 35 siswa SMA yang berpartisipasi dalam kompetisi tersebut diperiksa menggunakan model Rasch melalui program QUEST. Analisis mencakup kesesuaian butir (item fit), reliabilitas (reliabilitas butir dan person), tingkat kesukaran, dan efektivitas distraktor sebagai analisis

ARTICLE HISTORY
Received: February 20, 2026
Accepted: April 20, 2026

KEYWORDS:
Item Analysis, Quest Program, Rasch Model, Temu Fisika

KATA KUNCI:
Analisis Butir, Model Rasch, Program Quest, Temu Fisika

* Corresponding author:
Awal Mulia Rejeki Tumanggor, Physics Education Study Program, Universitas Negeri Manado, Tondano, Indonesia
✉ awaltumanggor@unima.ac.id ✆ +62 85-3714-92401

pelengkap berbasis CTT. Hasil menunjukkan bahwa 33 butir (82,5%) sesuai dengan model Rasch, sedangkan tujuh butir tidak sesuai (misfitting). Indeks reliabilitas Rasch tergolong baik (Reliabilitas Butir = 0,82, Reliabilitas Person = 0,84). Tingkat kesulitan terdistribusi dengan tepat, dengan 50% butir dikategorikan sedang, 22,5% mudah, 20% sulit, 2,5% sangat mudah, dan 5% sangat sulit. Namun, analisis distraktor mengungkapkan kekurangan yang signifikan: hanya 15% butir yang memiliki semua distraktornya berfungsi secara efektif. Studi ini merupakan analisis Rasch pertama dari olimpiade Temu Fisika, yang memberikan bukti empiris bahwa meskipun tes tersebut umumnya reliabel, perbaikan spesifik diperlukan. Temuan ini menyoroti pentingnya merevisi butir yang tidak sesuai dan distraktor yang tidak berfungsi untuk meningkatkan daya diskriminasi tes, khususnya di antara peserta dengan kemampuan tinggi. Studi ini juga memberikan referensi praktis untuk meningkatkan proses pengembangan dan validasi butir dalam kompetisi akademik.

A. Introduction

The success of an academic competition in the field of education is closely linked to the quality of its test items. One of the recurring academic events designed to evaluate students' competencies in physics is Temu Fisika, an annual program organized by the Department of Physics, Faculty of Mathematics, Natural Sciences, and Earth Sciences (FMIPAK), Universitas Negeri Manado. The competition aims to enhance students' quality, interest, and potential in achieving excellence and developing academic abilities in physics. Several studies have shown that participation in olympiad programs can significantly improve students' understanding of STEM concepts at the secondary school level (Tumanggor, et al., 2019; Smith, et al., 2021; Tschisgale, et al., 2024). These studies highlight that learning experiences gained through such competitions not only strengthen scientific knowledge and problem-solving abilities but also foster greater interest in science, technology, engineering, and mathematics. One of the main branches of Temu Fisika is the Senior High School Physics Olympiad for participants across North Sulawesi. The 2024 Physics Olympiad consisted of two sessions: Preliminary Round A and B, each comprising 40 multiple-choice questions presented in randomized order. The competition was participated in by selected students from various schools in North Sulawesi who had undergone internal school-based selection processes.

Physics olympiad questions are generally developed from fundamental concepts through analytical and contextual modifications, meaning that these problems should be solvable by students who have mastered the underlying principles but with a higher level of analytical complexity (Syarif & Syamsurizal, 2019). Thus, the quality of test items plays a crucial role in ensuring that the assessment instruments accurately measure students' competencies. According to the organizing committee of Temu Fisika 2024, the olympiad test items had undergone a qualitative analysis process covering content, construct, language, and key accuracy; however, a quantitative analysis had not yet been performed. Consequently, the statistical quality of each item remained undetermined, which could undermine the fairness and diagnostic value of

the competition. Without rigorous quantitative analysis, there is a risk that misfitting, poorly discriminating, or ineffective items may remain in use, potentially misrepresenting students' true abilities and reducing the validity of competition outcomes.

Quantitative item analysis serves as a valuable approach for evaluating test quality. While Classical Test Theory (CTT) has limitations in small-sample contexts, the Rasch model offers several advantages for analyzing competitive test items. First, it provides sample invariance, allowing item parameters to be estimated independently of the specific test-takers sample. Second, the Wright Map (person-item map) visually represents the alignment between item difficulty and student ability, revealing gaps in the measurement continuum. Third, it produces interval-scale measurement that enables more precise comparisons of ability and difficulty (Bond, et al., 2021; Boone, et al., 2014). These features make the Rasch model particularly suitable for high-stakes assessments such as academic competitions, where precise measurement and fairness are paramount.

Previous studies in other disciplines have demonstrated the utility of Rasch analysis in evaluating olympiad or competitive test items. For example, Rasch has been applied to assess biology olympiad questions (Syarif & Syamsurizal, 2019) and mathematics contest items (Hanna & Retnawati, 2022), revealing detailed insights into item functioning that CTT-based approaches often overlook. In physics education, Rasch analysis has been used to develop and validate diagnostic tests (Nirmala, et al., 2021; Tumanggor, et al., 2021), but its application to physics olympiad items, especially in the context of Temu Fisika, remains unexplored. To the best of the authors' knowledge, this study represents the first Rasch model analysis of physics olympiad items in the Temu Fisika competition. North Sulawesi has consistently participated in national physics olympiads, but there has been no systematic, empirical evaluation of the quality of olympiad test items used in local competitions such as Temu Fisika. Without such evaluation, item bias or poor distractor quality may go undetected, potentially disadvantaging students from this region when they compete at the national level. Therefore, this Rasch-based analysis provides a crucial foundation for improving the fairness and diagnostic validity of physics assessments in North Sulawesi, ultimately supporting more equitable preparation of students for national olympiads. Prior evaluations, if any, likely relied on CTT-based methods, which are more sensitive to sample characteristics and less robust for small populations. This gap underscores the novelty and necessity of the present study.

Therefore, this study aims to analyze the quality of senior high school physics olympiad test items used in the Temu Fisika 2024 competition by applying the Rasch model. The analysis includes evaluating item fit, reliability, difficulty level, discrimination power, and distractor effectiveness. The findings are expected to (1) provide a rigorous, sample-independent assessment of item quality; (2) identify

specific items that require revision or replacement; (3) contribute to the improvement of future Temu Fisika competitions and similar academic events; and (4) enrich the literature on Rasch-based evaluation of competitive assessments in STEM education.

B. Method

This study employed an evaluative research design with a quantitative approach to describe the measurement characteristics of the test items used in the competition. The purpose was not to test a hypothesis but to evaluate the quality of the physics olympiad test items from the Temu Fisika 2024 through the Rasch model. The research participants were 35 high school students who passed internal school selection and participated in the Preliminary Round Series A of the Temu Fisika 2024 Physics Olympiad in North Sulawesi. The research instrument consisted of the 40 multiple-choice items used in that round. Each item was scored dichotomously, with one point awarded for a correct answer and zero for an incorrect answer, without any penalty for guessing. Based on a detailed item-by-item classification, the test covered seven of the eleven major topics originally planned by the organizing committee. The distribution of items across the topics that actually appeared in the test is as follows: Translational Kinematics with Vector Analysis accounted for 8 items (20%), Translational Dynamics for 5 items (12.5%), Rotational Dynamics and Rigid Body Equilibrium for 1 item (2.5%), Fluid Statics and Dynamics for 6 items (15%), Heat and Thermodynamics for 10 items (25%), Electricity and Magnetism for 5 items (12.5%), and Modern Physics for 5 items (12.5%).

Data Collection Techniques

Data were collected using documentation techniques, specifically participants' answer sheets. Data analysis was performed using the QUEST software, which implements the Rasch 1-Parameter Logistic (1PL) model. The sample consisted of 35 participants from the Temu Fisika 2024 Physics Olympiad. Although the sample size is relatively small, Rasch analysis can still provide stable item parameter estimates provided the item reliability and separation index are at an adequate level (Wright & Stone, 1979; Linacre, 1994; Boone, et al., 2014; Bond, et al., 2021). However, a limited sample size may increase the standard error of parameter estimation, thereby reducing the precision of item and person measures. Linacre (1994) states that a sample of 30–50 participants can be used for initial (pilot) Rasch analysis, provided reliability and separation statistics are reported. Therefore, this study limits generalization and emphasizes the results as exploratory and evaluative for the context of this specific competition. The adequacy of the sample and the stability of the model estimates were further verified by examining the resulting fit statistics, reliability, and separation index.

Item Fit Analysis

The conformity of items to the model was measured using INFIT Mean Square (MNSQ) and INFIT t statistics. An item is considered fit if its INFIT MNSQ value falls within the range of 0.77 – 1.30 and its INFIT t value within the range of -2.0 to +2.0 (Sumintono & Widhiarso, 2015). Although wider thresholds (e.g., 0.5–1.5) are commonly suggested in the literature, a more stringent criterion was applied in this study to enhance the sensitivity of item fit detection, particularly in the context of a small sample and a high-stakes assessment setting. Items outside these ranges are declared misfit, indicating that the response pattern does not align with the measurement model.

Reliability and Separation

Within the Rasch model, reliability is reported as Person Reliability and Item Reliability, referring to the consistency and reproducibility of the measurement. Reliability values are interpreted as follows: ≥ 0.94 (excellent), 0.90–0.94 (very good), 0.80–0.90 (good), 0.67–0.80 (enough), and < 0.67 (weak). Values close to 1.0 indicate high consistency. The Separation Index is also reported, where a value > 2.0 is considered good for distinguishing between groups of participants or items with different ability/difficulty levels. This concept differs from Cronbach's Alpha in CTT and is more appropriate for evaluating internal consistency within the Rasch framework.

Item Difficulty

The item difficulty parameter (in logits) was estimated. The interpretation of difficulty level is based on the logit value, where large negative values indicate easy items and large positive values indicate difficult items (Boone, et al., 2014). For ease of interpretation, items were classified into five categories: very easy (< -2.0 logits), easy (-2.0 to -0.5 logits), moderate (-0.5 to +0.5 logits), difficult (+0.5 to +2.0 logits), and very difficult (> +2.0 logits) (Sumintono & Widhiarso, 2015). While alternative classification thresholds (e.g., ±0.86 logits) are commonly used, a broader categorization was applied in this study to allow finer interpretation of item spread, particularly in the context of a competitive test with varying levels of difficulty.

Distractor Effectiveness

As a complement to the Rasch analysis focusing on model parameters, an analysis of distractor (incorrect option) effectiveness was conducted using a CTT-based approach for diagnostic purposes. Distractor analysis was conducted by examining the frequency of selection for each incorrect option across participants. A distractor was declared functional if it was chosen by at least 5% of the participants (D'Sa & Visbal-Dionaldo, 2017). Distractors selected by fewer than 5% were

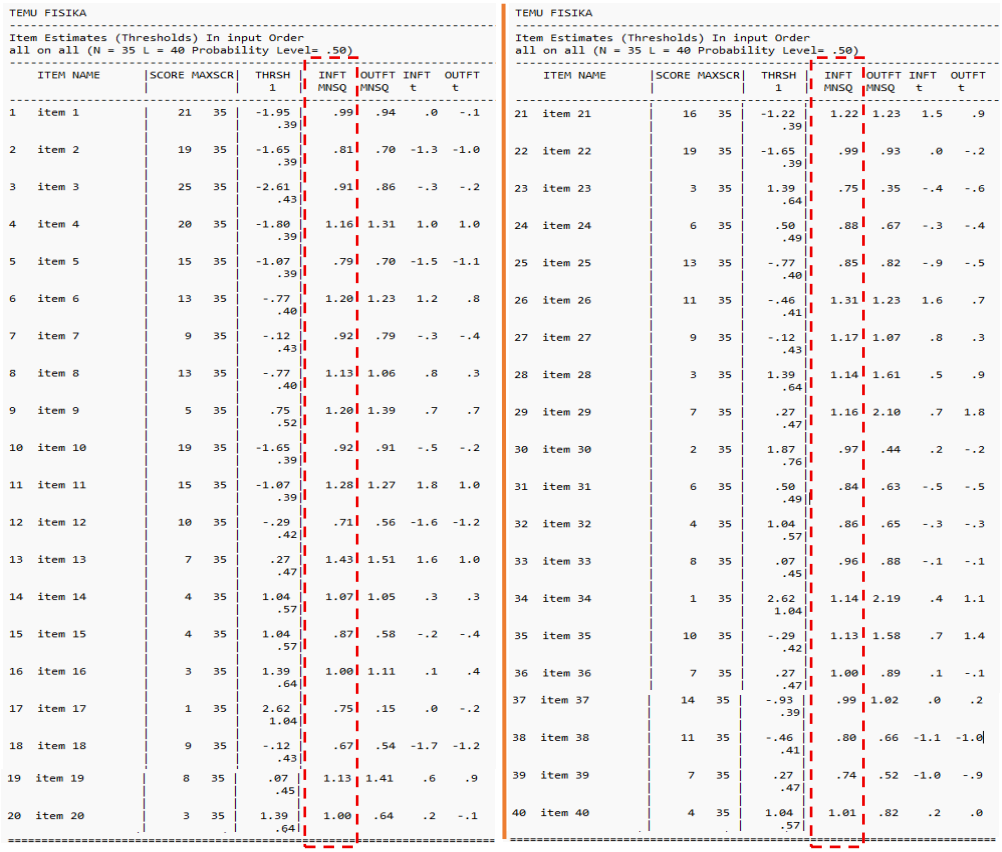
categorized as non-functional. This analysis helps identify implausible answer options or those that do not attract participants with lower conceptual understanding

C. Result and Discussion

The item analysis was conducted using dichotomous data, which were processed through the Rasch Model using the QUEST software. A total of 40 multiple-choice items from the Preliminary Round Series A of the Temu Fisika 2024 Physics Olympiad, answered by 35 participants, were analyzed. The Rasch Model was employed to evaluate response patterns and estimate parameters such as item fit, reliability, item difficulty, discrimination power, and distractor effectiveness.

Item Difficulty

The results of the item fit analysis based on the INFIT Mean Square (MNSQ) and INFIT t statistics showed that most of the items met the fit criteria in accordance with the Rasch Model. The item fit analysis was conducted using INFIT statistics obtained from the QUEST software. It should be noted that QUEST presents fit indices in a different format compared to other Rasch software; therefore, INFIT statistics were used as the primary indicator of item fit in this study.The results of analysis are presented in Figure 1.



TEMU FISIKA

Item Estimates (Thresholds) In input Order
all on all (N = 35 L = 40 Probability Level= .50)

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	21	35	-1.95 .39	.99	.94	.0	-.1
2	item 2	19	35	-1.65 .39	.81	.70	-1.3	-1.0
3	item 3	25	35	-2.61 .43	.91	.86	-.3	-.2
4	item 4	20	35	-1.80 .39	1.16	1.31	1.0	1.0
5	item 5	15	35	-1.07 .39	.79	.70	-1.5	-1.1
6	item 6	13	35	-.77 .40	1.20	1.23	1.2	.8
7	item 7	9	35	-.12 .43	.92	.79	-.3	-.4
8	item 8	13	35	-.77 .40	1.13	1.06	.8	.3
9	item 9	5	35	.75 .52	1.20	1.39	.7	.7
10	item 10	19	35	-1.65 .39	.92	.91	-.5	-.2
11	item 11	15	35	-1.07 .39	1.28	1.27	1.8	1.0
12	item 12	10	35	-.29 .42	.71	.56	-1.6	-1.2
13	item 13	7	35	.27 .47	1.43	1.51	1.6	1.0
14	item 14	4	35	1.04 .57	1.07	1.05	.3	.3
15	item 15	4	35	1.04 .57	.87	.58	-.2	-.4
16	item 16	3	35	1.39 .64	1.00	1.11	.1	.4
17	item 17	1	35	2.62 1.04	.75	.15	.0	-.2
18	item 18	9	35	-.12 .43	.67	.54	-1.7	-1.2
19	item 19	8	35	.07 .45	1.13	1.41	.6	.9
20	item 20	3	35	1.39 .64	1.00	.64	.2	-.1

TEMU FISIKA

Item Estimates (Thresholds) In input Order
all on all (N = 35 L = 40 Probability Level= .50)

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
21	item 21	16	35	-1.22 .39	1.22	1.23	1.5	.9
22	item 22	19	35	-1.65 .39	.99	.93	.0	-.2
23	item 23	3	35	1.39 .64	.75	.35	-.4	-.6
24	item 24	6	35	.50 .49	.88	.67	-.3	-.4
25	item 25	13	35	-.77 .40	.85	.82	-.9	-.5
26	item 26	11	35	-.46 .41	1.31	1.23	1.6	.7
27	item 27	9	35	-.12 .43	1.17	1.07	.8	.3
28	item 28	3	35	1.39 .64	1.14	1.61	.5	.9
29	item 29	7	35	.27 .47	1.16	2.10	.7	1.8
30	item 30	2	35	1.87 .76	.97	.44	.2	-.2
31	item 31	6	35	.50 .49	.84	.63	-.5	-.5
32	item 32	4	35	1.04 .57	.86	.65	-.3	-.3
33	item 33	8	35	.07 .45	.96	.88	-.1	-.1
34	item 34	1	35	2.62 1.04	1.14	2.19	.4	1.1
35	item 35	10	35	-.29 .42	1.13	1.58	.7	1.4
36	item 36	7	35	.27 .47	1.00	.89	.1	-.1
37	item 37	14	35	-.93 .39	.99	1.02	.0	.2
38	item 38	11	35	-.46 .41	.80	.66	-1.1	-1.0
39	item 39	7	35	.27 .47	.74	.52	-1.0	-.9
40	item 40	4	35	1.04 .57	1.01	.82	.2	.0

Figure 1. Analysis Results of Each Item

Based on the analysis results in Figure 1, specifically, items 13 and 26 exceeded the upper threshold of the INFIT MNSQ range (>1.30), while items 12, 17, 18, 23, and 39 fell below the lower limit (<0.77). Therefore, these seven items were categorized as misfit and may not adequately represent the intended construct. Although seven items were identified as misfitting, a detailed qualitative content analysis was not conducted in this study. Such analysis is important to identify potential sources of misfit, including conceptual errors, ambiguous wording, or poorly functioning distractors. To strengthen the argument for construct validity in the Rasch analysis, Item 17 (Figure 2) was selected as a representative example of a misfitting item. This item was intended to assess students' understanding of fluid statics through the application of Archimedes' principle, particularly the relationship between an object's density and the fraction of its volume submerged in water. However, the qualitative analysis suggests that the observed statistical misfit may be influenced by construct-irrelevant variance. Specifically, the phrase "the height of each material in the water" lacks a clear operational definition, allowing multiple interpretations that can lead to different solution strategies among students with comparable physics understanding. In addition, several distractors reflect numerical permutations rather

than meaningful physics-based misconceptions, thereby emphasizing numerical pattern recognition or computational accuracy over conceptual reasoning. Together, these factors likely contributed to response patterns that are inconsistent with the assumptions of the Rasch model, resulting in the observed item misfit.

Three materials with masses m_1, m_2, and m_3 have densities of $\frac{1}{2}\rho_{\text{water}}$, ρ_{water}, and $2\rho_{\text{water}}$, respectively. These three materials are placed into a cylindrical container filled with water. The cylinder has a radius of 50 cm and a surface area of 2.01×10^2 m. The height of each material in the water is … m, … m, and … m ($\rho_{\text{water}} = 1000\ \text{kg/m}^3$).

A. 500, 1000, and 2000
B. 63.51, 31.76, and 0 *
C. 31.76, 63.51, and 127.02
D. 71.62, 51.36, and 227.01
E. 127.02, 63.51, and 31.76

Figure 2. Item 17 on static fluid material

In contrast, the remaining items met the fit requirements, suggesting that they conformed to the model expectations. The visualization of the item fit map shown in Figure 2 confirmed that most items were well distributed within the acceptable range between 0.77 and 1.30, as indicated in previous studies by Abdulfattah & Supahar, (2019) and Bond, et al., (2021). This alignment indicates that the majority of the test items function appropriately in measuring the participants' ability continuum. Items falling outside this range should be revised or replaced in future implementations to improve the overall quality and construct validity of the assessment.

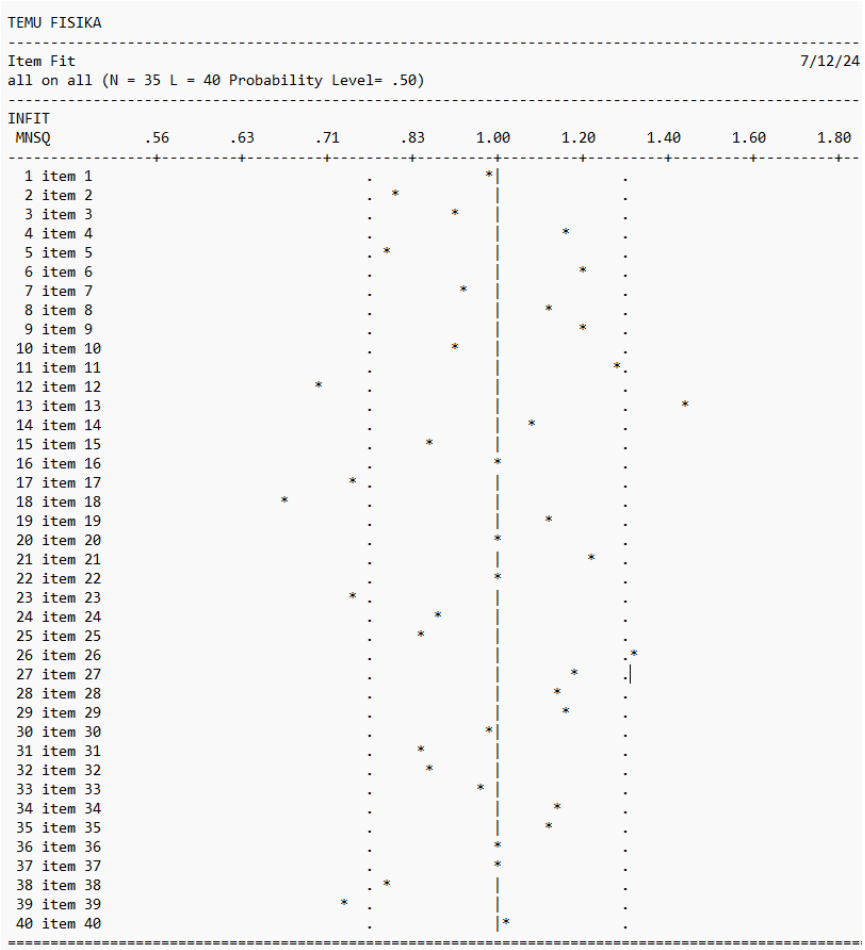


Figure 3. Fit Map of Rasch Model

Reliability Estimation

The reliability analysis using the Rasch Model produced satisfactory results. As shown in Table 1 of the original dataset, the reliability coefficient for item estimates was 0.82, while the reliability coefficient for person estimates was 0.84.

Table 1. Results of Item and Case Estimation

Test Parameter	Item Estimates	Case Estimates
Mean & Std Dev.	0.00 ± 1.24	-1.38 ± 1.14
Adjusted Std Dev.	1.12	1.05
Mean & Std Dev. of INFIT MNSQ	1.00 ± 0.18	1.00 ± 0.15
Mean & Std Dev. of OUTFIT MNSQ	0.98 ± 0.44	0.98 ± 0.42
Mean & Std Dev. of INFIT t	0.07 ± 0.87	0.06 ± 0.68
Mean & Std Dev. of OUTFIT t	0.07 ± 0.76	0.07 ± 0.70
Reliability	0.82	0.84

According to the reliability classification by George & Mallery (2020), both coefficients fall into the "good" category, meaning that the test instrument demonstrates consistent measurement performance. The high reliability value for the item estimates indicates that the items as a whole possess a strong internal consistency in measuring the intended construct. Meanwhile, the high reliability value for person estimates implies that participants' responses were stable and consistent, suggesting serious engagement with the test items.

These findings collectively confirm that the physics olympiad test used in Temu Fisika 2024 is reliable and appropriate for evaluating students' physics competence in an academic competition context. Furthermore, this aligns with Azwar (2015) and Hanna & Retnawati (2022), who emphasized that reliability reflects the precision and dependability of an instrument in producing consistent data.

Item Difficulty Estimation

The difficulty levels of the test items were analyzed using the item threshold estimates generated by the Rasch analysis in the QUEST software. The summary of item difficulty levels is presented in Table 2.

Table 2. Recapitulation of Item Difficulty

No	Threshold	Criteria	No	Threshold	Criteria
1	1.95	Easy	21	-1.22	Easy
2	-1.65	Easy	22	-1.65	Easy
3	-2.61	Very Easy	23	1.39	Difficult
4	-1.80	Easy	24	0.50	Medium
5	-1.07	Easy	25	-0.77	Medium
6	-0.77	Medium	26	-0.46	Medium
7	-0.12	Medium	27	-0.12	Medium
8	-0.77	Medium	28	-1.39	Easy
9	-0.75	Medium	29	0.27	Medium
10	-1.65	Easy	30	1.87	Difficult
11	-1.07	Easy	31	0.50	Medium
12	-0,29	Medium	32	1.04	Difficult
13	0.27	Medium	33	0.07	Medium
14	1.04	Difficult	34	2.62	Very Difficult
15	1.04	Difficult	35	-0.29	Medium
16	1.39	Difficult	36	0.27	Medium
17	2.62	Very Difficult	37	-0.93	Medium
18	-0.12	Medium	38	-0.46	Medium
19	0.07	Medium	39	0.27	Medium
20	1.39	Difficult	40	1.04	Difficult

Based on the analysis results, one item (2.5%) was classified as very easy, nine items (22.5%) were easy, twenty items (50%) were moderate, eight items (20%) were difficult, and two items (5%) were very difficult. This distribution of item difficulty approximates a normal curve, which indicates that the test items are well balanced

across various levels of challenge. Such a distribution is considered ideal in educational assessment, as it allows the test to effectively differentiate participants with different levels of physics competence.

In the context of learning evaluation, items that are too easy do not provide sufficient challenge to stimulate critical thinking or higher-order reasoning. Conversely, items that are excessively difficult may discourage participants and distort the measurement of true ability. Therefore, the composition of difficulty levels observed in this study suggests that the test has achieved a proportional balance, ensuring that the physics olympiad assessment is both fair and diagnostically meaningful (Arikunto, 2015; Tumanggor & Supahar, 2021).

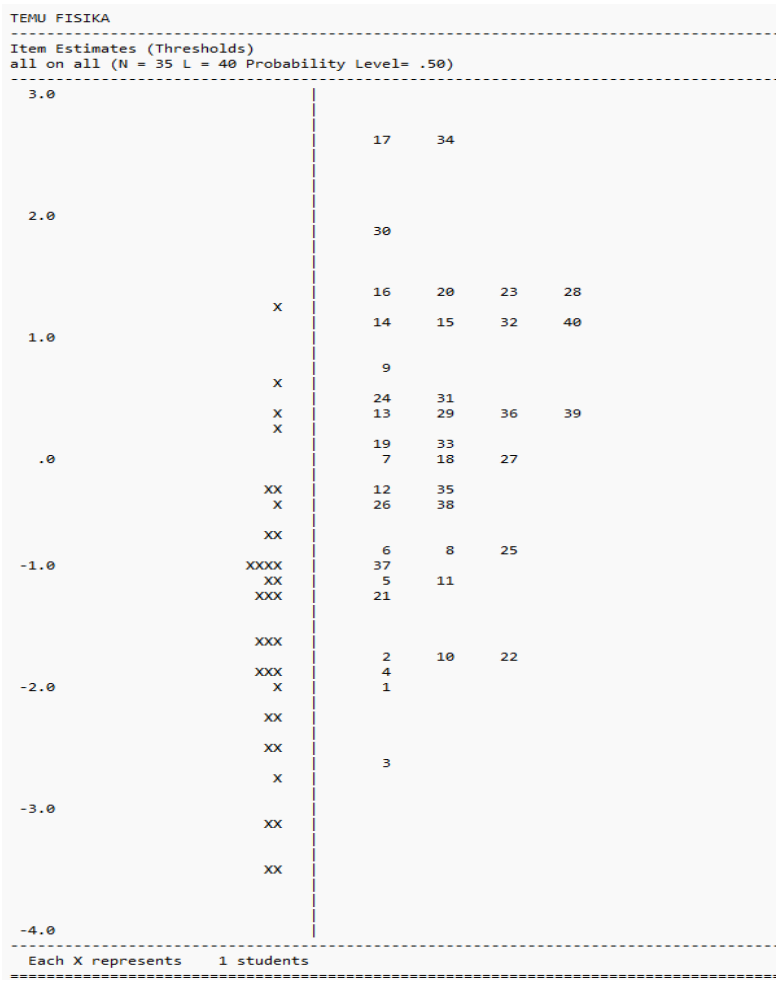


Figure 4. Wright Map of Rasch Model

The Wright Map (item-person map) presented in Figure 4 reveals a clear gap at the higher ability levels, particularly in the logit range above +2.0. Within this range, only two items (Items 17 and 34) are present, indicating that the instrument does not

provide a sufficient number of items to adequately target high-ability participants. This condition suggests that the test lacks measurement precision at the upper end of the ability continuum, limiting its capacity to distinguish between highly proficient and exceptionally proficient students, which is a critical objective in an Olympiad context. Consequently, this gap reduces the sensitivity of the instrument in measuring extreme high-level abilities and indicates that future test development should incorporate more high-difficulty items to improve construct coverage and strengthen measurement validity at the upper ability range (Tumanggor & Supahar, 2020).

Distractor Effectiveness

As a complementary analysis based on CTT, the effectiveness of distractors (incorrect options) was examined. The analysis of distractor performance revealed that only 15% of the items had distractors functioning effectively, whereas the remaining 85% contained at least one or more non-functional distractors. More specifically, 42.5% of the items had one non-functional distractor, 35% had two, and 7.5% contained three ineffective distractors. This result indicates that a substantial proportion of distractors were not plausible or sufficiently aligned with common misconceptions. Effective distractors are essential in multiple-choice assessments because they increase the discriminative power of an item by attracting participants who have not yet mastered the relevant concept (Ceuppens, et al., 2018).

Items with poorly functioning distractors tend to produce less reliable measurement outcomes. For instance, if a distractor is chosen by fewer than 5% of participants, it can be considered non-functional and should be revised to improve its plausibility Istiyono (2018). In some cases, even conceptually strong items may require revision to optimize distractor quality. Additionally, items that produce error results are considered invalid and should not be used in subsequent testing rounds, as they may compromise the overall reliability and validity of the instrument (Kamilah & Suwarna, 2016; Tumanggor, et al., 2020).

Although the complementary analysis based on Classical Test Theory (CTT) revealed that most items contained poorly functioning distractors, the overall test reliability remained in the good category. At first glance, this finding appears paradoxical, as weak distractors theoretically increase guessability and may undermine measurement quality. However, several factors may account for this outcome. First, within the context of a physics Olympiad, participants typically exhibit relatively high ability and homogeneous reasoning patterns, enabling them to readily eliminate poorly constructed distractors and identify the correct options. As a result, the presence of weak distractors may not substantially disrupt score consistency across examinees, thereby sustaining a high reliability coefficient. Second, the relatively small sample size (N = 35) may also contribute to stable or even inflated reliability estimates, particularly when variability in participant ability is limited. Consequently, the high reliability observed in this study should not be interpreted as

evidence that all aspects of item quality are optimal, but rather as a reflection of participant characteristics and test context. These findings underscore the importance of interpreting reliability with caution and complementing it with distractor analysis and Rasch fit statistics to obtain a more comprehensive evaluation of instrument quality.

D. Conclusion

This study presents an evaluation of the quality of senior high school Physics Olympiad test items used in the Temu Fisika 2024 competition through the application of the Rasch model, complemented by a Classical Test Theory–based analysis. Overall, the findings indicate that most items are capable of measuring students' abilities consistently; however, several limitations remain, particularly with respect to misfitting items, suboptimal distractor performance, and gaps in item difficulty at the upper end of the ability continuum. These results underscore that instrument quality is not determined solely by overall reliability, but also by item fit, coverage of difficulty levels, and clarity of item construction. From a practical perspective, the findings provide a clear foundation for the development of a Physics Olympiad item bank. Items that meet the Rasch model fit criteria may be retained as core items, whereas misfitting items and those with ineffective distractors should be revised or removed. Furthermore, the Wright Map analysis highlights the need to include more high-difficulty items to enhance measurement precision in distinguishing students with very high levels of ability, which is a primary objective in academic competitions. This approach supports a more systematic and sustainable process of item calibration for future Olympiad administrations. Overall, this study contributes to the field of science education evaluation and academic competition assessment by demonstrating that the Rasch model offers a robust and informative framework for examining the quality of Physics Olympiad instruments. The findings are expected to inform item developers and assessment researchers in improving the validity, precision, and sustainability of science competition assessments at the secondary school level.

References

Abdulfattah, A., & Supahar. (2019). The development of high school physics problem solving skills test instruments based problem-based learning. *Journal for the Education of Gifted Young Scientists*, *7*(4), 1037–1052. https://doi.org/10.17478/jegys.602291

Arikunto, S. (2015). *Dasar-dasar Evaluasi Pendidikan*. Bumi Aksara.

Azwar, S. (2015). *Reliability and Validity* (fourth). Pustaka Pelajar.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model, Fundamental Measurement in the Human Sciences*. Routledge. https://doi.org/10.4324/9781410614575

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in The Human Sciences*. Springer.

Ceuppens, S., Deprez, J., Dehaene, W., & De Cock, M. (2018). Design and validation of a test for representational fluency of 9th grade students in physics and mathematics: The case of linear functions. *Physical Review Physics Education Research*, *14*(2), 20105. https://doi.org/10.1103/PhysRevPhysEducRes.14.020105

D'Sa, J. L., & Visbal-Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, *9*(3), 109. https://doi.org/10.5958/0974-9357.2017.00079.4

George, D., & Mallery, P. (2020). *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference* (Sixteenth). Routledge. https://doi.org/10.4324/9780429056765

Hanna, W. F., & Retnawati, H. (2022). Analisis Kualitas Butir Soal Matematika Menggunakan Model Rasch Dengan Bantuan Software Quest. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *11*(4), 3695. https://doi.org/10.24127/ajpm.v11i4.5908

Istiyono, E. (2018). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika dengan Teori Tes Klasik dan Modern*. UNY Press.

Kamilah, D. S., & Suwarna, I. P. (2016). Pengembangan Three-Tier Test Digital Untuk Mengidentifikasi Miskonsepsi Pada Konsep Fluida Statis. *Edusains*, *8*(2), 212–220. https://doi.org/10.15408/es.v8i2.5192

Linacre, J. M. (1994). Sample Size and Item Calibration or Person Measure Stability. *Rasch Measurement Transactions*, *7*(4), 328.

Nirmala, M. F. T., Tumanggor, A. M. R., & Supahar. (2021). Analysis of Validity and Reliability of Diagnostic Test of Picture Representation Ability in High School Physics Learning. *Proceedings of the 7th International Conference on Research, Implementation, and Education of Mathematics and Sciences (ICRIEMS 2020)*, *528*(Icriems 2020), 538–546. https://doi.org/10.2991/assehr.k.210305.079

Smith, K. N., Jaeger, A. J., & Thomas, D. (2021). "Science Olympiad Is Why I'm Here": The Influence of an Early STEM Program on College and Major Choice. *Research in Science Education*, *51*, 443–459. https://doi.org/10.1007/s11165-019-09897-7

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata Publishing House.

Syarif, E. A., & Syamsurizal, S. (2019). Analyzed Quality of Senior High School Biology Olympiad Questions at West Sumatera, Riau, Jambi, and Bengkulu in 2018. *Bioeducation Journal*, *3*(2), 142–150. https://doi.org/10.24036/bioedu.v3i2.194

Tschisgale, P. L., Steegh, A., Petersen, S., Kubsch, M., Wulff, P., & Neumann, K. (2024). Are science competitions meeting their intentions? A case study on affective and cognitive predictors of success in the Physics Olympiad. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1). https://doi.org/10.1186/s43031-024-00102-y

Tumanggor, A. M. R., Jumadi, J., Wilujeng, I., & Ringo, E. S. (2019). The Profile of Students' Physics Problem Solving Ability in Optical Instruments. *Jurnal*

Penelitian & Pengembangan Pendidikan Fisika, *5*(1), 29–40. https://doi.org/10.21009/1.05104

Tumanggor, A. M. R., & Supahar. (2021). Dissemination of Online Diagnostic Test Instrument to Diagnose High School Students' Mathematical Representation Ability: The Case of Work and Energy. *Proceedings of the 6th International Seminar on Science Education (ISSE 2020)*, *541*, 528–537. https://doi.org/10.2991/assehr.k.210326.076

Tumanggor, A. M. R., Supahar, & Nirmala, M. F. T. (2021). The Development of Diagnostic Test Instrument for Verbal Representation Ability in High School Physics Learning. *Proceedings of the 7th International Conference on Research, Implementation, and Education of Mathematics and Sciences (ICRIEMS 2020)*, *528*, 471–476. https://doi.org/10.2991/assehr.k.210305.069

Tumanggor, A. M. R., & Supahar, S. (2020). The development of diagnostic test instrument for mathematical representation ability (PhysDTRA) in high school physics learning. *Journal for the Education of Gifted Young Scientists*, *8*(4), 1439–1453. https://doi.org/10.17478/jegys.777425

Tumanggor, A. M. R., Supahar, S., Ringo, E. S., & Harliadi, M. D. (2020). Detecting Students' Misconception in Simple Harmonic Motion Concepts Using Four-Tier Diagnostic Test Instruments. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, *9*(1), 21–31. https://doi.org/10.24042/jipfalbiruni.v9i1.4571

Wright, B. D., & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Dalam *Mesa Press*. Routledge.