JIPF (JURNAL ILMU PENDIDIKAN FISIKA)
p-ISSN: 2477-5959 | e-ISSN: 2477-8451 *Vol. 9 No. 3, September 2024, Page 474-486*

Item Quality Analysis of Physics Concept Understanding Test with Rasch Model

Aris Kurniawan [1*)], Edi Istiyono [2], Sul Daeng Naba [3]
Universitas Negeri Yogyakarta, Indonesia[1,2,3]
[*)]Corresponding E-mail: ariskurniawan.2023@student.uny.ac.id

Received: June 19th, 2024. Revised: August 11th, 2024. Accepted: August 13th, 2024

Keywords :
Concept Understanding; Rasch Model; Item Response Theory

ABSTRACT
This study aims to determine the quality of physics concept understanding test items on the topic of electromagnetic waves based on estimation of item fit, item difficulty level estimation, and item reliability estimation. The research sample consisted of 298 eleventh-grade science students from several public high schools in Central Kalimantan province. The results of data analysis using the Rasch model with the aid of QUEST software showed that all items were fit or valid, with INFIT MNSQ values ranging from 0.77 to 1.33. The item reliability was 0.98, categorized as excellent, and the test participant reliability was 0.25, categorized as weak. The difficulty level of the test items was categorized into three: 5 items were difficult, 5 items were moderate, and 5 items were easy. Furthermore, based on the analysis using OUTFIT t values, one item was discarded and could not be used as a test instrument because it had an OUTFIT t value greater than 2.

INTRODUCTION

Evaluation and assessment in the learning process are crucial components that determine the effectiveness of learning and the success of students [1]. Good assessment can deliver reliable information about students' competency achievements and the effectiveness of the teaching methods used [2]. One way to ensure accurate assessment is through comprehensive item analysis [3]. Item trials are used to produce high-quality test items, ensuring that the measurements conducted using these items are more accurate and reliable.

Students' ability to understand concepts is one of the skills commonly measured in education, especially in the field of physics education. Understanding physics concepts plays a crucial role in students' learning at the secondary school level. Concepts in physics, such as force, energy, motion, waves, and light, form the foundational understanding needed to comprehend natural phenomena and modern technological applications [4]. According to Pennington & Black [4], strong understanding of physics concepts enables students to relate theories to physical phenomena observed in everyday life. This aligns with Land [5], which show that a good grasp of physics concepts is related to students' ability to solve physics problems effectively.

Understanding concepts is a crucial skill for students and must be mastered. By understanding concepts, students can expand their learning abilities and apply the concepts learned to real-life situations. Students' ability to understand concepts can be measured by their ability to comprehend or understand something from various perspectives [6]. To enhance students' conceptual understanding, it is necessary to not only develop the learning process but also support the development of concept understanding assessments through tests designed to meet standards that encompass content, construction, and language requirements, and possess high validity and reliability.

There are several indicators of conceptual understanding. These indicators consist of seven aspects: interpreting, explaining, summarizing, exemplinifing, classifying, inferring, and comparing [7]. According to Riwanto et al [8], the indicators of conceptual understanding include classifying, explaining, interpreting, comparing, and exemplinifing. Anderson & Krathwohl [9] suggest that the indicators of conceptual understanding encompass generalizing, inferring, interpreting, comparing, exemplinifing, explaining, and classifying. In this study, the indicators of conceptual understanding measured in the test instrument include interpreting, classifying, exemplinifing, comparing, and explaining, as these align with cognitive competencies and can build conceptual knowledge.

The electromagnetic wave spectrum is one of the complex topics in the physics curriculum for XI grade science students in high school. Understanding this concept requires an in-depth comprehension of various physical phenomena related to electromagnetic waves, making it essential to assess students' understanding meticulously and carefully [10]. To ensure that the test items used to evaluate students' understanding of this material are of high quality, item trials are conducted.

Item Response Theory (IRT) has proven to be an effective tool in item analysis as it can provide more detailed and accurate information compared to classical methods [11]. IRT is a theoretical framework used to design, analyze, and assess the quality of tests and assessment instruments. In contrast to classical theory that focuses on total scores, IRT focuses on the relationship between individual abilities and question item characteristics [12]. IRT is based on several important assumptions, they are unidimensionality, local independence, and parameter invariance [13]. Rasch model is one popular model in IRT, which provides a strong framework for analyzing item characteristics. The model assumes that the probability of a correct answer depends only on the difference between individual ability and item difficulty [14] [15]. The Rasch model is one of the simplest and most frequently used IRT models. In the context of grain quality analysis, the model provides important information such as item difficulty and item match to the model [3].

Some of the previous studies that used the Rasch model were conducted by Glamočić et al [16] using analysis of Differential Item Function (DIF) to evaluate bias in items in physics tests. Their findings suggest that DIF analysis can identify items that may be biased against certain groups, such as gender or ethnic background, and allow for the development of fairer and more valid tests. Wahyuni et al [17] develop and validate measurement instruments for computational thinking skills among physics education students. The study found that the use of the Rasch model in data analysis provided results that showed high reliability and validity, supporting the use of this instrument in the evaluation of physics education. Research conducted by Planinic et al [18] analyzed the Force Concept Inventory (FCI) using the Rasch. The results show that Rasch's model can identify non-compliant items and provide insights into how students understand the concept of style, which is crucial for the improvement of assessment instruments.

Syadiah & Hamdu [19] utilized the Rasch model to analyze critical thinking test questions in STEM learning. The results of the study showed suitability to identify the quality of test items or students, thus providing satisfaction to teachers in carrying out the learning process at school. Nisa et al [20] demonstrate that the Rasch can be used to test the validity and reliability of test items, including analyzing the DIF to detect gender and domicile bias. This aligns with the research conducted by Tarigan et al [21], which shows that the use of the Rasch provides good accuracy in determining the validity and feasibility of test instruments.

Based on various studies, the Rasch model has proven to be highly effective in estimating students' abilities, evaluating item fit, and determining the reliability of test items. It offers more accurate information than traditional methods and helps identify and correct weaknesses in evaluation instruments. The Rasch model can be used to test the validity and reliability of an instrument, assess item and respondent reliability, check dimensionality, and detect bias in items [22] [23] [24]. It also determines student abilities, where a student's likelihood of answering an item correctly correlates with their ability level [25] [26]. In essence, the Rasch model can analyze both the instrument items and the students themselves [27] [28]. In this study, the Rasch model was used to analyze a concept understanding test instrument on electromagnetic wave material. The Rasch model was chosen for analyzing students' conceptual understanding tests because of its ability to provide more objective and reliable measurements. This model can separate item difficulty from student ability, offering a more accurate representation of students' conceptual understanding [23]. Additionally, the Rasch model excels in diagnostic analysis, as evidenced by the use of Wright maps, which visualize both student abilities and item difficulties on the same scale, facilitating the identification of comprehension gaps [2] [22].

Conceptual understanding is a crucial skill that students must develop during the learning process. A strong conceptual understanding allows students to build a more complex cognitive structure, making it easier to connect one concept with another [29] [30]. Therefore, it is essential to have a reliable measurement tool for assessing students' conceptual understanding. A test instrument is considered effective if it is valid, reliable, and usable [31] [32] [33] [34], enabling accurate measurement of students' abilities [35] [36]. Instrument validity can be established through content, construct, and criterion-related validation processes [31] [34]. In this study, the Rasch model is employed to analyze the quality of physics concept understanding test items in electromagnetic wave material, focusing on the estimation of item fit, item difficulty level, and item reliability.

Previous research conducted by Hofer et al [37] used the Rasch model to analyze a conceptual understanding test instrument on Newton's mechanics. The findings indicated that the Rasch model produced a fair, rigorous, and efficient measurement tool for assessing high school students' conceptual understanding of Newtonian mechanics. Similarly, research by Bozdağ & Türkoğuz [38] demonstrated the effectiveness of the Rasch model in analyzing elementary students' conceptual understanding of light, confirming that the test was fully valid and reliable. Handayani et al [35] further supported these findings by showing that the Rasch model effectively analyzed the reliability and validity of a conceptual understanding test on electricity and magnetism, making it a robust tool for measuring students' understanding. Andriani et al [39] also highlighted the Rasch model's ability to evaluate the validity, reliability, and difficulty level of test items in a basic physics course. Lastly, Fiskawarni et al [36] utilized the Rasch model to validate 20 test items and scoring rubrics for measuring energy literacy among prospective physics teachers, confirming their suitability for use.

Numerous studies have demonstrated the efficacy of the Rasch model in accurately estimating student abilities, evaluating item fit, and enhancing test reliability compared to traditional methods [35] [36] [37] [38] [39]. By identifying poorly functioning items, the Rasch model offers a robust approach to improving assessment quality. Despite its advantages, the application of the Rasch model in physics education, particularly in regions like Central Kalimantan, Indonesia, remains limited. The results of field observations suggest that most public high schools in this province, including SMAN 1 Balai Riam, SMAN 1 Permata Kecubung, SMAN 2 Kasongan, SMAN 1 Sukamara, SMAN 1 Montalat, and SMAN 1 Palangka Raya, have yet to adopt rigorous item analysis procedures. This lack of attention to assessment quality raises concerns about the validity and reliability of student performance data.

To address this gap, objectives of this study to analyze the quality of physics concept understanding test items related to electromagnetic waves using the Rasch model. By employing the QUEST software, a well-established tool for Rasch analysis, this research will provide valuable insights into item fit, difficulty, and item reliability. The research focuses on three main objectives: (1) calibrating

test items, (2) analyzing the fit of each test item, and (3) assessing the overall reliability and validity of the test. The findings are expected to contribute to the development of more effective assessment practices in physics education and inform instructional improvements in Central Kalimantan. Additionally, this study will contribute to provide a set of validated test items to assess students' understanding of electromagnetic waves, as well as demonstrating the usefulness of the Rasch model and QUEST software in physics education research.

Based on the description, research is needed to address the following research questions, including: (1) How is the estimated fit of the concept understanding test item on electromagnetic wave material to the Rasch model? (2) How is the item difficulty level estimation of the concept understanding test item on electromagnetic wave materials? (3) What is the reliability of the concept understanding test item on electromagnetic wave material?

METHOD

This study is a descriptive research with a quantitative approach to assess the quality of the test instrument. This research will be carried out in May 2024. The population of this study is all class XI students in public high schools in Central Kalimantan province. The sample in this study consisted of 298 eleventh-grade science students. In determining the sample, the researcher used the Quota sampling technique. The sample taken by the researcher is students from several schools representing three locations, namely sub-districts, districts, and urban areas. These schools are SMAN 1 Balai Riam, SMAN 1 Montalat, SMAN 1 Permata Kecubung as schools located in the sub-district, SMAN 2 Kasongan and SMAN 1 Sukamara as schools located in the district, and SMAN 1 Palangka Raya as schools located in urban areas. The stages conducted in this research were: 1) determining the subject matter; 2) preparing the item indicators of the instrument; 3) composing the instrument items; 4) validating the content by experts; 5) conducting instrument trials with test participants; and 6) analyzing the results of the instrument trials [40]. The instrument used is a concept comprehension test consisting of 15 multiple-choice items on electromagnetic wave material. Data collection was done using an online-based Google Form. Data analysis technique employed the Rasch model using the QUEST program. Data analysis was conducted to assess item quality through five stages: 1) estimation of item fit; 2) item difficulty level estimation; 3) reliability estimation; and 4) estimation of passed items.

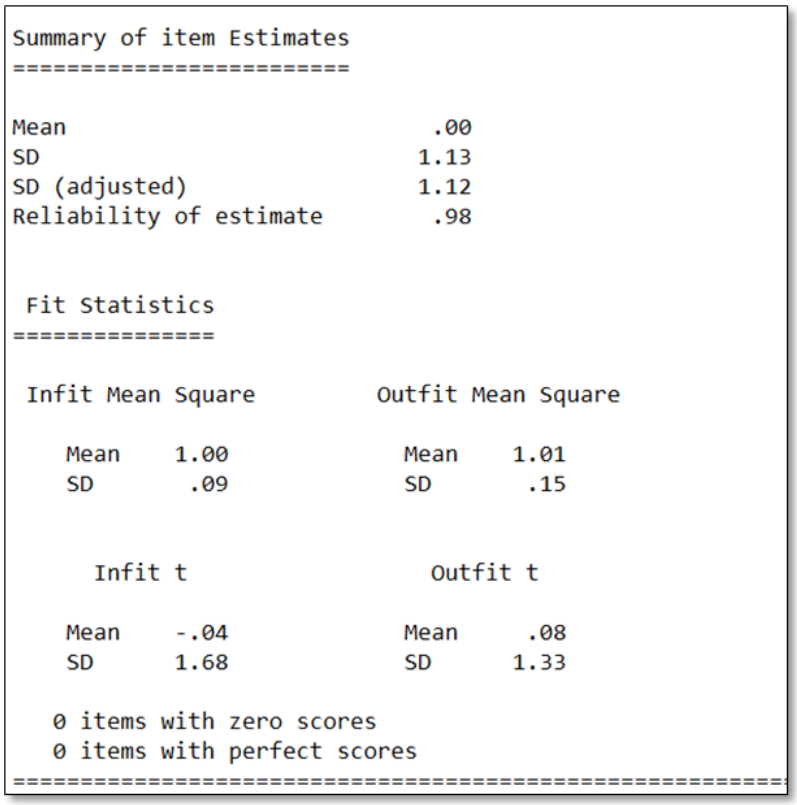
RESULTS AND DISCUSSIONS

The results of this study include: estimation of item fit, item difficulty level estimation, item reliability estimation, and estimation of item passed.

Estimation of Item Fit

As stated by Setyawarno [41], the overall fit of items with the Rasch model is evaluated based on the average value of the INFIT Mean Square (INFIT MNSQ) and the standard deviation. The fit of each specific item with the Rasch model is determined by the OUTFIT t value of the item in question. The results of fit statistics is shown in Figure 1.

Based on Figure 1, the average INFIT MNSQ value of 1.00 with a standard deviation of 0.09 is obtained. Further analysis will an INFIT MNSQ range of 1.00 ± 0.09, resulting in a range from 0.91 to 1.09. Meanwhile, the OUTFIT t value is 0.08 ± 1.33, resulting in a range from -1.25 to 1.41. Based on the results of the data analysis, the INFIT MNSQ value range is 0.91-1.09 (which falls within the acceptable range of 0.77-1.33), and the OUTFIT t value is ≤ 2, which if interpreted using the INFIT MNSQ and OUTFIT t value criteria presented in Table 1 and Table 2 [41] [42] [43] It can be inferred that, overall, all items fit with the Rasch model.



```
Summary of item Estimates
=========================

Mean                          .00
SD                           1.13
SD (adjusted)                1.12
Reliability of estimate       .98


 Fit Statistics
===============

 Infit Mean Square         Outfit Mean Square

    Mean    1.00              Mean    1.01
    SD       .09              SD       .15


      Infit t                   Outfit t

    Mean    -.04              Mean     .08
    SD      1.68              SD      1.33

   0 items with zero scores
   0 items with perfect scores
============================================================
```

Fig 1. *Recapitulation of Item Fit Statistics*

Table 1. Value Criteria of INFIT MNSQ

Value	Criteria
> 1.33	Does not fit the Rasch model
0.77 – 1.33	Fit the Rasch model
< 0.77	Does not fit the Rasch model

Table 2. Value Criteria of OUTFIT MNSQ

Value	Criteria
≤ 2	Fit the Rasch model
> 2	Does not fit the Rasch model

Estimation of item fit can also be said to be a test of the validity of an item. Validity refers to the alignment of a test with what should be measured, in other words, the test instrument used is accurate in measuring the predetermined variable [44]. An item has high validity if it can measure the expected competence. However, if an item cannot measure the expected competence, then it has low validity. Items with high validity are able to measure students' abilities. Conversely, items with low validity need improvement [45]. The conformity of items with the Rasch Model can be assessed using the INFIT MNSQ values. Setyawarno [41] stated that the INFIT MNSQ value can be used to determine how well each item fits with the specified model criteria. The recapitulation of INFIT MNSQ values for each item analyzed through the QUEST program is shown in Figure 2.

Table 1 indicates that items fit with the Rasch model, and meeting the requirements, have INFIT MNSQ values ranging from 0.77 to 1.33. Based on the data analysis as seen in Figure 2, all items in position within the range from 0.77 to 1.33, indicating that all items used are fit with the Rasch model or valid. This analysis result indicates that all items fit or meet the criteria for INFIT MNSQ values, thus it can be concluded that the items are appropriate for assessing students' conceptual understanding in electromagnetic wave material.

QUEST: The Interactive Test Analysis System

Item Estimates (Thresholds) In input Order
all on all (N = 298 L = 15 Probability Level= .50)

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	54	298	1.62 .16	1.06	1.24	.6	1.6
2	item 2	149	298	.08 .12	.88	.87	-3.4	-1.9
3	item 3	241	298	-1.41 .15	.97	.94	-.3	-.4
4	item 4	220	298	-1.00 .14	.92	.87	-1.2	-1.2
5	item 5	79	298	1.13 .14	1.03	1.13	.5	1.2
6	item 6	68	298	1.33 .14	.99	.96	-.1	-.3
7	item 7	82	298	1.08 .14	1.00	1.00	.0	.1
8	item 8	193	298	-.56 .13	1.11	1.14	2.2	1.6
9	item 9	238	298	-1.34 .15	.85	.80	-1.8	-1.5
10	item 10	233	298	-1.24 .15	.91	.81	-1.2	-1.6
11	item 11	158	298	-.05 .12	1.00	1.00	.1	.1
12	item 12	57	298	1.56 .15	1.02	1.03	.2	.3
13	item 13	156	298	-.02 .12	.96	.96	-1.1	-.6
14	item 14	239	298	-1.36 .15	1.13	1.27	1.5	1.9
15	item 15	141	298	.19 .12	1.13	1.16	3.6	2.1
Mean				.00	1.00	1.01	.0	.1
SD				1.13	.09	.15	1.7	1.3

Fig 2. *Recapitulation of Item Estimate*

Besides using the INFIT MNSQ values as seen in Figure 2, the item fit map presented in Figure 3 can also be used to determine whether the items used are relevant or fit with the Rasch model.

Test items that fit the Rasch model are within the range from 0.77 to 1.33 [41] [42] or are within the area between the two dashed lines in the figure. Based on Figure 3, out of the 15 items tested, all are evenly distributed within this range or between the two dashed lines. Therefore, it can be inferred that all tested items fit the Rasch model or are valid. The validity test of test items is used to map the items with various criteria. Good test items as assessment tools must be valid, reliable, and usable [35] [36]. In this study, the student conceptual understanding test on electromagnetic waves was developed as a

multiple-choice test with five answer options. Before item analysis using the Rasch model, the test instrument underwent content validation by four experts using a validation sheet, followed by revisions based on their feedback. The content-validated instrument was then tested on 298 eleventh-grade science students in public high schools in Central Kalimantan province. The results were analyzed using the Rasch model with QUEST software to estimate item fit, item difficulty level, and item reliability. The analysis showed that all tested items fit the Rasch model and were valid. Validity and reliability testing of test instruments is crucial to ensure accurate and trustworthy measurement results [35]. Test items are considered to have high validity when the item score is parallel to the total score [30] [46] [47]. Arifin [48] stated that several factors influence the validity of test items, including participants' responses, scoring of participants' answers, the test instrument used, and evaluation administration. Factors related to participants' responses include the tendency of participants to provide spontaneous answers and difficulties in recognizing scientific language in the tested items [30].

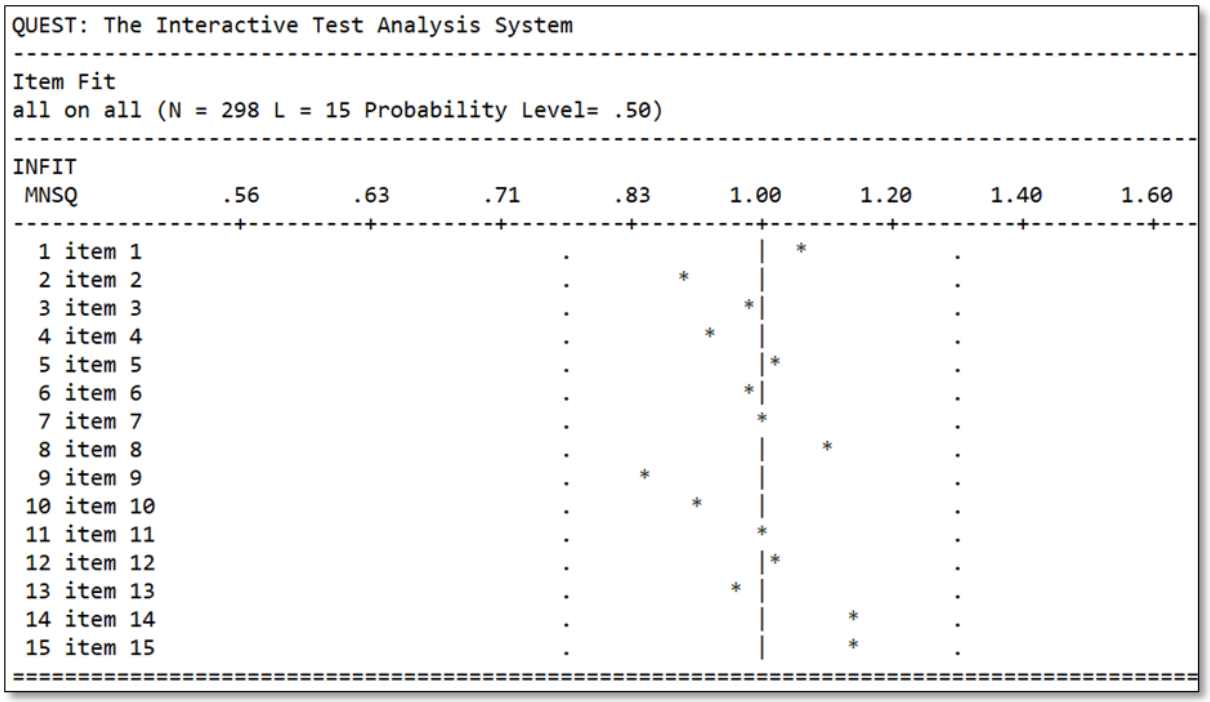


Fig 3. *Fit map of Rasch model*

Item Difficulty Level Estimation

The difficulty level of items is used to distinguish the ability of test participants who can answer correctly from those who cannot [30]. The difficulty level of items analyzed using the item estimate value (Threshold). The difficulty level criteria range from -2.0 to 2.0. An item is considered very easy if it is less than or equal to -2.0, and very difficult if it is greater than 2.0. The threshold value criteria and the distribution of the difficulty levels of the test items are presented in Table 3 [41] [43] and Table 4.

Table 3. Criteria of Threshold Value

Value of Threshold	Criteria
$b > 2$	Very Difficult
$1 < b \leq 2$	Difficult
$-1 < b \leq 1$	Moderate
$-2 < b \leq -1$	Easy
$b \leq -2$	Very Easy

Table 4. Item Difficulty Level Estimation

Item	Threshold Value	Criteria
1	1.62	Difficult
2	0.08	Moderate
3	-1.41	Easy
4	-1.00	Easy
5	1.13	Difficult
6	1.33	Difficult
7	1.08	Difficult
8	-0.56	Moderate
9	-1.34	Easy
10	-1.24	Easy
11	-0.05	Moderate
12	1.56	Difficult
13	-0.02	Moderate
14	-1.36	Easy
15	0.19	Moderate

Table 4 indicates that the difficulty levels of items are divided into several categories: 1) items 1, 5, 6, 7, and 12 fall into the DIFFICULT category, 2) items 2, 8, 11, 13, and 15 fall into the MODERATE category, and 3) items 3, 4, 9, 10, and 14 fall into the EASY category. If presented in percentage form, out of the 15 test items administered to 298 test participants, 33.33% of the items are categorized as DIFFICULT, 33.33% as MODERATE, and 33.33% as EASY. Ideally, test items should have a balanced proportion of difficulty levels, with 25% of items in the easy category, 50% in the moderate category, and 25% in the difficult category [30]. Furthermore, high-quality test items are those that are neither too easy nor too difficult. This is because questions that are too easy may not stimulate enough effort to answer, and questions that are too difficult may cause frustration as they are beyond the test participants' abilities [30] [49].

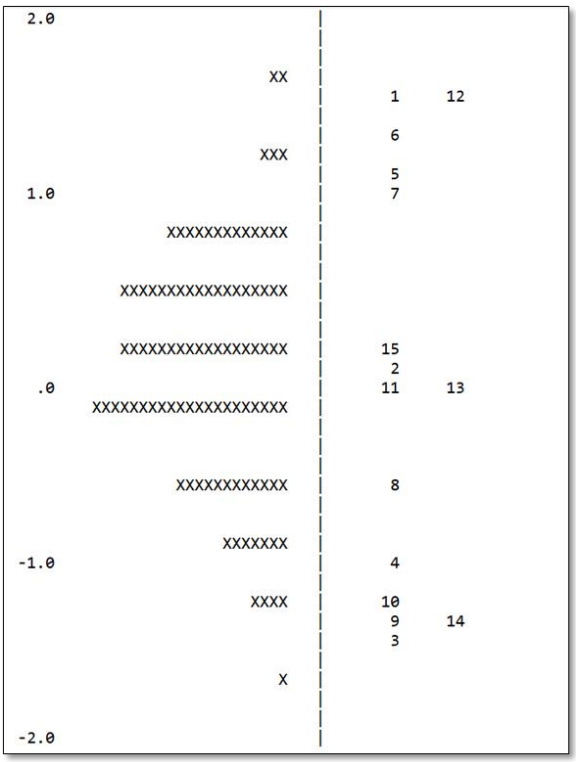


Fig 4. *Wright Map*

In addition to using the value of the item estimate (threshold), the difficulty level of items can also be estimated from the Wright Map. Wright Map is used to visualize the distribution of test participants' abilities and item abilities, which helps in analyzing the difficulty level of items based on the logit scale [50] [51]. By observing the position of each item on the Wright Map, then comparing it to the logical scale, we can determine which items fall into the difficult, medium, and easy categories. The distribution of the difficulty level of items is shown in Figure 4.

Based on Figure 4, the estimation of the difficulty level of the item can be done by observing the position of points 1 to 15 on the Wright Map which are distributed to the left of the dotted line, while for the "x" sign distributed to the left of the dotted line is the distribution of the ability of test participants. The Wright Map presents the distribution of test participants' abilities according to the difficulty level of the item on a logit scale of -2.0 to +2.0. From the Wright Map, it can be estimated that item 1 and item 12 are the most difficult items, because their position is close to the logit scale +2.0, while item 3 is the easiest item, because their position is close to the logit scale -2.0. Then, if an analysis is carried out to estimate the level of ability of test participants, then based on the distribution of test participants' abilities on the Wright Map, a normal curve will be produced (see Figure 5), which means that there are few test participants with low and high abilities, while test participants with moderate ability are the most numerous [41].

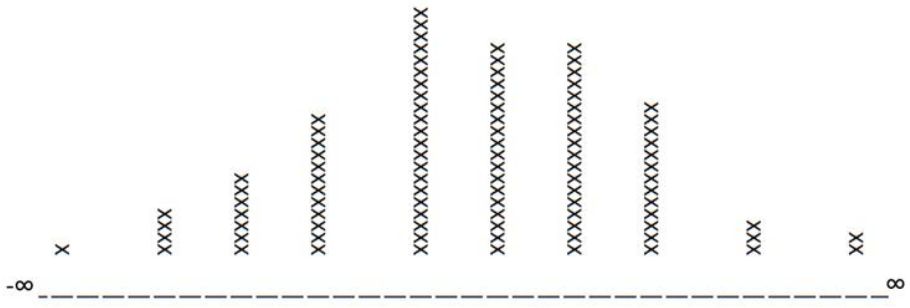


Fig 5. *Person Ability Level*

Estimation of item reliability

Reliability refers to the consistency of an instrument. A test is considered reliable if it yields consistent results when administered to the same group at different times [30]. The reliability of items is determined by item reliability and person reliability values using the Rasch model [2] [22]. The Rasch model was employed to assess the reliability of items based on the item reliability estimate, and it can also be used to assess the ability of test participants based on the reliability of case estimate. The reliability criteria for item estimate and case estimate in the Rasch model are shown in Table 5 [22].

Table 5. Reliability Criteria in the Rasch Model

Value of Case and Item Estimate Reliability	Criteria
> 0.94	Excellent
0.91 – 0.94	Very Good
0.81 – 0.90	Good
0.67 – 0.80	Moderate
< 0.67	Weak

Based on the data analysis, the reliability of items and the test participants is presented in Table 6. The reliability of items is 0.98, which is categorized as excellent. This shown that the quality of the test items used to measure students' understanding of the electromagnetic wave spectrum is highly consistent. Meanwhile, the reliability of the test participants is 0.25, categorized as weak [43]. This suggests that there is inconsistency among the test participants in answering the test items [30] [52]

[53]. In other words, the consistency of the test participants in answering questions about the electromagnetic wave spectrum is weak or inconsistent.

Tabel 6. Item and Test Participant Reliability

	Value of Case and Item Estimate Reliability	Criteria
Item Reliability	0.98	Excellent
Test Participants	0.25	Weak

If the value of the case estimate reliability is in the good category, it indicates that the test participants' responses are consistent [53]. Research conducted by Pratama [53] only 67 test participants were used and a reliability of case estimate value of 0.38 was obtained with a weak category, while the research conducted by Purba [54] to analyze the achievement test instrument which amounted to 50 items using the Rasch model was 428 students. The item reliability value obtained is 0.99, and the person reliability value is 0.92. This shows that the number of test participants affects the reliability score of test participants. The more test participants, the higher the reliability value [53]. This statement is reinforced by Hakiki et al [55] findings with a sample of 417 students, the results of the reliability test participants were obtained of 0.76 and item reliability of 0.97. Furthermore, the findings by Istiyono et al [56] with a sample of 1001 students and the number of items used as many as 44 items, the reliability value of the instrument was obtained of 0.95.

Estimation of Passed Items

The OUTFIT t value was employed to assess which items fail or pass. Items pass or succeed if the OUTFIT t value is less than equal to 2.00, and fail if the OUTFIT t value is more than equal to 2.00 [41] [43]. A recapitulation of the OUTFIT t values obtained from Figure 1 is presented in Table 7.

Table 7. Fit Item Recapitulation

Item	OUTFIT t Value	Description
1	1.6	Passed
2	-1.9	Passed
3	-0.4	Passed
4	-1.2	Passed
5	1.2	Passed
6	-0.3	Passed
7	0.1	Passed
8	1.6	Passed
9	-1.5	Passed
10	-1.6	Passed
11	0.1	Passed
12	0.3	Passed
13	-0.6	Passed
14	1.9	Passed
15	2.1	Passed

Table 7 shows that items 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14 have an OUTFIT t value of less than 2.00. Therefore, it can be concluded that these 14 items have passed, meaning they can be used as test instruments to measure students' conceptual understanding of electromagnetic wave material. Meanwhile, item 15 has an OUTFIT t value greater than 2.00, so it is considered invalid and cannot be used.

CONCLUSION AND SUGGESTION

The concept understanding test instrument on electromagnetic wave material was developed and tested

on 298 XI grade MIPA students from several high schools in Central Kalimantan province. The analysis was then conducted using the Rasch model. The results were categorized as having good quality based on validity indicated by the INFIT MNSQ values, where it was concluded that all items were fit or valid with values ranging from 0.77 to 1.33. The item reliability analysis showed a value of 0.98, categorized as excellent, while the test-taker reliability showed a value of 0.25, categorized as weak. The item difficulty level, based on the item estimate (threshold) values, was categorized into three: 5 items with high difficulty, 5 items with moderate difficulty, and 5 items with low difficulty. Additionally, based on the analysis using the OUTFIT t values, item 15 was found to be invalid and could not be used as a test instrument because it had an OUTFIT t value greater than 2.

REFERENCES

[1] Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc..
[2] Popham, W. J. (2008). *Classroom assessment: What teachers need to know*. Pearson.
[3] Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences, Third Edition*. Psychology Press.
[4] Pennington, M. C., & Black, S. L. (2010). Conceptual Understanding and Scientific Reasoning of High School Students in Physics. *International Journal of Science Education*, *32*(5), 567–589.
[5] Land, T. (2013). Conceptual understanding: The case of electricity and magnetism. *European Journal of Science and Mathematics Education*, *1*(1), 13–28.
[6] Bella, A. Z., Azizahwati, A., & Azhar, A. Pengembangan Instrumen Tes Pemahaman Konsep Materi Momentum dan Impuls. *Jurnal Online Mahasiswa (JOM) Bidang Keguruan dan Ilmu Pendidikan*, *8*(1), 117-127.
[7] Abdi, M. U., Mustafa, M., & Pada, A. U. T. (2021). Penerapan pendekatan STEM berbasis simulasi PhET untuk meningkatkan pemahaman konsep fisika peserta didik. *JIPI (Jurnal IPA dan Pembelajaran IPA)*, *5*(3), 209-218.
[8] Riwanto, D., Azis, A., & Arafah, K. (2019). Analisis pemahaman konsep peserta didik dalam menyelesaikan soal-soal fisika kelas x mia sma negeri 3 soppeng. *Jurnal Sains Dan Pendidikan Fisika*, *15*(2), 23-31.
[9] Afifah, R. (2019). Analisis profil proses kognitif pemahaman konsep siswa. *Jurnal Pendidikan Fisika*, *7*(2), 170-178.
[10] Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Wadsworth Publishing.
[11] Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Prentice Hall.
[12] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.
[13] Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC press.
[14] Rash, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
[15] Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
[16] Salibašić Glamočić, D., Mešić, V., Neumann, K., Sušac, A., Boone, W. J., Aviani, I., ... & Grubelnik, V. (2021). Maintaining item banks with the Rasch model: An example from wave optics. *Physical Review Physics Education Research*, *17*(1), 010105.
[17] Purnami, W., Fauzi, A., & Naingalis, M. L. P. (2023, June). Computational thinking skills identification among students of physics education department using Rasch model analysis. In *AIP Conference Proceedings* (Vol. 2751, No. 1). AIP Publishing.

[18] Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics—Physics Education Research*, *6*(1), 010103.
[19] Syadiah, A. N., & Hamdu, G. (2020). Analisis rasch untuk soal tes berpikir kritis pada pembelajaran STEM di sekolah dasar. *Premiere Educandum: Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(2), 138-148.
[20] Nisa, K., & Suprapto, N. (2023). Deteksi Bias Gender dan Domisili Menggunakan DIF (Differential Item Functioning): Analisis Instrumen Tes Keterampilan Pemecahan Masalah Terintegrasi Etnofisika. *Inovasi Pendidikan Fisika*, *12*(1), 30-35.
[21] Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). Analisis Instrumen Tes Menggunakan Rasch Model dan Software SPSS 22.0. *Jurnal Inovasi Pendidikan Kimia*, *16*(2), 92-96.
[22] Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.
[23] Maulana, S., Rusilowati, A., Nugroho, S. E., & Susilaningsih, E. (2023, June). Implementasi rasch model dalam pengembangan instrumen tes diagnostik. In *Prosiding Seminar Nasional Pascasarjana* (Vol. 6, No. 1, pp. 748-757).
[24] Suryani, Y. E. (2018). Aplikasi rasch model dalam mengevaluasi Intelligenz Structure Test (IST). *Psikohumaniora: Jurnal Penelitian Psikologi*, *3*(1), 73-100.
[25] Islam, A. A., Gu, X., Crook, C., & Spector, J. M. (2020). Assessment of ICT in tertiary education applying structural equation modeling and Rasch model. *Sage Open*, *10*(4), 2158244020975409.
[26] Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research*, *14*(2), 020104.
[27] Matore, M. E. E. M., Maat, S. M., Affandi, H. M., & Khairani, A. Z. (2018). Assessment of psychometric properties for Raven Advanced Progressive Matrices in measuring intellectual quotient (IQ) using Rasch model. *Asian Journal of Scientific Research*, *11*(3), 393-400.
[28] bin Khairani, A. Z., & bin Abd Razak, N. (2015). Modeling a multiple choice mathematics test with the Rasch model. *Indian Journal of Science and Technology*, *8*(12), 1.
[29] Ozarslan, M., & Çetin, G. (2018). Biology Students' Cognitive Structures about Basic Components of Living Organisms. *Science Education International*, *29*(2).
[30] Wulandari, T., Ramli, M., & Muzzazinah, M. (2022). Analisis butir soal dynamic assessment untuk mengukur pemahaman konsep klasifikasi tumbuhan pada mahasiswa. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, *10*(1), 191-201.
[31] Groundlund, R. L., & Linn, N. E. (1990). *Measurement and Evaluation in Teaching*. Prentice Hall College Div.
[32] Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test*. Routledge Taylor & Francis Group.
[33] Engeström, Y., Virkkunen, J., Helle, M., Pihlaja, J., & Poikela, R. (1996). The change laboratory as a tool for transforming work. *Lifelong learning in Europe*, *1*(2), 10-17.
[34] Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology, Fourth Ed*. Wadsworth/Thomson Learning.
[35] Handayani, Y., Rahmawati, R., & Widiasih, W. (2023). Using Rasch Model to Analyze Reliability and Validity of Concept Mastery Test on Electricity and Magnetism Topic. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *8*(2), 226-239.
[36] Fiskawarni, T. H., Rahmawati, R., Widiasih, W., & Saad, R. (2024). The Design and Validation of the Four Tier Test Instrument for Energy Literacy Using the Rasch Model Analysis. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *9*(1), 74-87.
[37] Hofer, S. I., Schumacher, R., & Rubin, H. (2017). The test of basic Mechanics Conceptual Understanding (bMCU): using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton's mechanics. *International journal of STEM education*, *4*, 1-20.
[38] Bozdağ, H. C., & Türkoğuz, S. (2021). A Rasch Model Analysis of Primary School Students'

Conceptual Understanding Levels of the Concept of Light. *International Online Journal of Primary Education*, *10*(1), 160-179.

[39] Andriani, R., Fadieny, N., & Permana, N. D. (2023). Development of Conceptual Understanding Student Tests to The Basic Physics Subject: a Rasch Model Analysis. *Co-Catalyst: Journal of Science Education Research and Theories*, *1*(1), 43-54.

[40] Retnawati, H. (2016). *Analisis Kuantitatif Instrumen Penelitian (Panduan Peneliti, Mahasiswa, dan Psikometrian)*. Prama Publishing.

[41] Setyawarno, D. (2017). Penggunaan AplikasiSoftware Iteman (Item and Test Analysis) untuk Analisis Butir Soal Pilihan Ganda Berdasarkan Teori Tes Klasik. *Jurnal Ilmu Fisika Dan Pembelajarannya (JIFP)*, *1*(1), 11-21.

[42] Subali, B., & Suyata, P. (2011). *Panduan analisis data pengukuran pendidikan untuk memperoleh bukti empirik kesahihan menggunakan program Quest*. Lembaga Penelitian dan Pengabdian Masyarakat UNY: Yogyakarta.

[43] 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), 3695-3704.

[44] Ghazali, N. H. M. (2016). A Reliability and Validity of an Instrument to Evaluate the School-Based Assessment System: A Pilot Study. *International journal of evaluation and research in education*, *5*(2), 148-157.

[45] Setyaningrum, P. M. P., Ramli, M., & Rinanto, Y. (2018). Analisis kualitas butir soal instrumen assessment diagnostic untuk mendeteksi miskonsepsi siswa SMA pada materi virus. *Jurnal Bioedukatika*, *6*(2), 91-101.

[46] Borualogo, I. S., Kusdiyati, S. K., Susandari, S. S., & Sirodj, D. A. N. (2017). Analisis item soal UTS pedologi semester ganjil 2015-2016. *Schema: Journal of Psychological Research*, 46-57.

[47] Nofiana, M. (2017). Pengembangan instrumen evaluasi higher orderthinking skills pada materi kingdom plantae. *Pedagogi Hayati*, *1*(1).

[48] Arifin, Z. (2011). *Evaluasi Pembelajaran*. Remaja Rosdakarya.

[49] Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12-23.

[50] Khatmani, F. K., Liliawati, W., & Imansyah, H. Analisis pemahaman konsep fisika siswa smp pada materi suhu dan kalor menggunakan tes isomorfik: Rasch model. *WaPFi (Wahana Pendidikan Fisika)*, *9*(1), 11-18.

[51] Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 020111.

[52] Ardiyanti, D. (2016). Aplikasi model rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karir siswa. *Jurnal Psikologi*, *43*(3), 248-263.

[53] Pratama, D. (2020). Analisis kualitas tes buatan guru melalui pendekatan item response theory (IRT) model rasch. *Tarbawy: Jurnal Pendidikan Islam*, *7*(1), 61-70.

[54] Purba, S. E. D. (2018). Analisis model Rasch instrumen tes prestasi pada mata pelajaran dasar dan pengukuran listrik. *Wiyata Dharma: Jurnal Penelitian Dan Evaluasi Pendidikan*, *6*(2), 142-147.

[55] Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analisis properti psikometri subtes merkaufgaben (ME) dengan rasch model. *Jurnal Psikologi*, *14*(1), 40-49.

[56] Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA. *Jurnal penelitian dan evaluasi pendidikan*, *18*(1), 1-12.