KPJ 8(2) (2024)
KAPPA JOURNAL
Physics & Physics Education
https://e-journal.hamzanwadi.ac.id/index.php/kpj/index

The Application of Rasch Model to Analyse the Validity and Reliability of an Instrument for Reflective Thinking Skills on Topic of Wave-Particle Dualism

Tarpin Juandi[1,4], Ida Kaniawati[1], Achmad Samsudin[2], Lala Septem Riza[3]
[1]Pendidikan Ilmu Pengetahuan Alam, Universitas Pendidikan Indonesia, Indonesia
[2]Pendidikan Fisika, Universitas Pendidikan Indonesia, Indonesia
[3]Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia, Indonesia
[4]Pendidikan Fisika, Universitas Hamzanwadi, Indonesia

Received: 18 July 2024
Revised: 22 July 2024
Accepted: 30 August 2024

Corresponding Author:
Ida Kaniawati
kaniawati@upi.edu

DOI:
https://doi.org/10.29408/kpj.v8i2.27049

Abstract: This study aims to analyse the validity and reliability of an instrument for assessing reflective thinking skills on the topic of wave-particle dualism in modern physics lectures using the Rasch model. The Rasch model was selected for its capability to provide a more in-depth analysis of item performance and respondent ability, as well as to identify misfitting or biased items. The research method employed is a descriptive quantitative approach, utilizing Winsteps software for data analysis. The sample consists of 36 students enrolled in modern physics lectures at a university in West Nusa Tenggara. The results indicate that the instrument has excellent item reliability (0.91) and excellent internal consistency (Cronbach's Alpha 0.86), although the respondent reliability falls into the weak category (0.62). The instrument's validity also meets the Rasch model's acceptance criteria, with infit MNSQ and outfit MNSQ values ranging from 0.5 to 1.5. Further analysis reveals that some items are misfitting and need revision to ensure fairness and consistency in measuring reflective thinking skills. These findings make a significant contribution to the development of more accurate and reliable assessment tools in physics education

Keywords: Modern Physics, Rasch Model, Reflective Thinking, Reliability and Validity

Introduction

Technological advancements in the field of psychometrics, such as the Rasch model, offer a more comprehensive approach to measuring the validity and reliability of educational evaluation instruments. The Rasch model, which is part of item response theory, allows for a more in-depth analysis of item performance and respondent ability (Ling Lee et al., 2021; Samsudin et al., 2021; Sumintono and Widhiarso, 2015). It provides a more objective and transparent framework for assessing the quality of an instrument, ensuring that it truly measures what it is intended to measure. Modern physics requires evaluation instruments that can assess not only lower-order thinking skills but also higher-order thinking skills in students. High-order thinking instruments are crucial to ensure that the learning process can facilitate the development of high-order thinking skills, which are essential for understanding complex physics concepts (Arabatzis, 2017; Kanim, 2020; Syahidi et al., 2023). In this context, valid and reliable evaluation becomes urgent to identify and enhance the quality of learning and students' understanding of the subject matter.

Many existing instruments are often not thoroughly tested for validity and reliability, resulting in evaluation outcomes that may be inaccurate and unreliable. Additionally, limitations in traditional analytical methods often fail to reveal in-depth

How to Cite:
Juandi, T, Kaniawati, I., Samsudin, A., & Riza, L. S. (2024). The Application of Rasch Model to Analyse the Validity and Reliability of an Instrument for Reflective Thinking Skills on Topic of Wave-Particle Dualism. *Kappa Journal, 8*(2), 270-277. https://doi.org/10.29408/kpj.v8i2.27049

information about item performance and respondent ability (Wei et al., 2020; Widhiarso and Sumintono, 2016). The main issue in this research is the lack of valid and reliable evaluation instruments to measure reflective thinking skills in modern physics lectures. This is particularly true for topics such as black body radiation, X-ray Compton effect, photoelectric effect, and the Bohr model for the hydrogen atom (Alves and Santos, 2021; Sun and Latora, 2020). Obtaining a reliable instrument is indeed challenging; however, a common solution applied in various studies is the use of the Rasch model to analyze response data. The Rasch model allows researchers to perform more comprehensive and detailed analyses, including the ability to identify misfitting or biased items and to assess the internal consistency of the instrument (Planinic et al., 2019; Samsudin et al., 2021). Therefore, the use of the Rasch model can improve the accuracy and reliability of evaluation results, as well as provide better insights into how the instrument functions in different contexts.

The use of the Rasch model has been extensively applied to evaluate and improve measurement instruments in the field of education. Research by Boone and Staver (2020) applied the Rasch model in the context of evaluating higher-order thinking skills. Their study results showed that the Rasch model could identify biased items and improve the internal consistency of the instrument, ensuring that the instrument could be used more effectively to measure higher-order thinking skills. They also emphasized the importance of individual item analysis to understand how each item contributes to the overall construct of the instrument. Another study by Bond and Fox (2015) demonstrated that the Rasch model could be effectively used to assess the validity and reliability of instruments in a more objective manner compared to classical methods. In this study, the Rasch model enabled the identification of poorly functioning items and provided useful information for instrument revision. In the context of physics education, a study by Juandi et al. (2023) applied the Rasch model to evaluate instruments used in modern physics lectures. The findings of this research indicated that the Rasch model not only enhanced the validity and reliability of the instruments but also provided deeper insights into students' understanding of complex physics concepts. This helps in identifying areas that require improvement in the curriculum and teaching methods.

Although the Rasch model has been widely used in various educational contexts, there is still a gap in research applying this model to measure reflective thinking skills, particularly in modern physics lectures. Many previous studies have focused more on evaluating basic cognitive knowledge without considering the dimension of higher-order thinking skills (Maryani et al., 2021; Nitriani et al., 2022; Setiawan et al., 2021). Additionally, many instruments used in research have not undergone rigorous validation processes using the Rasch model, resulting in potentially less accurate and reliable outcomes. Furthermore, most studies have focused on either validity or reliability, without integrating both comprehensively (Hadzhikolev et al., 2020; Salido and Dasari, 2019). This highlights the need for more in-depth and holistic research that not only evaluates but also improves measurement instruments for reflective thinking in the context of modern physics lectures.

The novelty of this research lies in its comprehensive application of the Rasch model, which not only assesses but also enhances the quality of evaluation instruments in the context of physics education. The scope of this research includes collecting data from students attending modern physics lectures and analysing the data using the Rasch model. The results of this research can make a significant contribution to the field of physics education by providing more accurate and reliable evaluation tools to measure students' reflective thinking skills. This study aims to analyse valid and reliable instruments for measuring reflective thinking skills in modern physics lectures using the Rasch model.

Method

A descriptive quantitative method was used in this research with the application of Winsteps in Rasch modeling for data analysis. Rasch analysis was chosen for its ability to provide a more comprehensive assessment of item performance and respondent ability, as well as to identify and correct misfitting or biased items (Boone and Staver, 2020). The sample in this study consisted of 36 students attending modern physics lectures at a university in West Nusa Tenggara. The sample was randomly selected to ensure data representativeness. Students were asked to complete a test consisting of items designed to measure reflective thinking skills. The test was administered offline to ensure accuracy in data collection. The instrument consisted of several sections covering various aspects of reflective thinking, such as analysis, evaluation, inference, and reflection (Laliyo et al., 2022). The collected data were then analysed using Winsteps software for the Rasch model.

Before data collection, the instrument was developed based on relevant literature and validated by experts to ensure that it was easy to understand and relevant to the context of modern physics lectures. This pilot testing provided feedback that was used to revise

and improve the instrument, ensuring that the final instrument used for data collection from a larger sample was more valid and reliable. The parameters measured in this study included item difficulty level, respondent ability, and internal consistency of the instrument. Item difficulty was measured based on how many students answered the item correctly, while respondent ability was measured based on their performance on the entire instrument. The internal consistency of the instrument was measured using reliability coefficients generated from Rasch model analysis. Additionally, analyses were conducted to identify biased or misfitting items that needed revision (Bond and Fox, 2015; Handayani et al., 2023; Rouquette et al., 2019).

Results and Discussion

This section outlines the instrument-level validity, item-level validity, and instrument reliability. Instrument-level validity includes data fit with the Rasch model and construct validity. Meanwhile, item-level validity encompasses how much and how precise, how good, and differential item functioning (DIF).

1. Instrument-Level Validity

In the analysis using the Rasch model, instrument validity is determined based on instrument-level validity and item-level validity. Instrument-level validity includes construct validity and data fit with the model (whether the data fits the model). Meanwhile, item-level validity encompasses how good, Differential Item Functioning (DIF), and how much and how precise. The following are explanations of each.

a. Construct Validity

Construct validity is assessed by the values of raw variance explained by measures, unexplained variance in the 1st and 2nd contrasts, and eigenvalues. These variable values can be seen in Table 1.

Table 1. Standardized Residual Variance in Eigenvalue Units and Item Information Units

Variabel	Eigenvalue	Observed	Expected
Raw Variance Explained by Measures	16.5148	50.8%	51.0%
Raw Variance Explained by Persons	2.5088	7.7%	7.7%
Raw Variance Explained by Items	14.0061	43.1%	43.3%
Unexplained Variance in 1st Contrasts	2.8713	8.8%	17.9%
Unexplained Variance in 2nd Contrasts	2.2901	7.0%	14.3%

The raw variance explained by measures (the combination of persons and items) is 16.5148 or 50.8%, which is very close to the expected value of 50.8%. This indicates that the Rasch model adequately explains the variance in the data. The raw variance explained by persons (differences among individuals) is 2.5088 or 7.7%. This value matches the expected value of 7.7%, indicating that individual differences contribute significantly to the data variance. The raw variance explained by items is 14.0061 or 43.1%, which is very close to the expected value of 43.3%. This indicates that differences in item difficulty contribute significantly to the data variance.

In Rasch model analysis, there are factors that cannot be explained (unexplained variance) but contribute to the data variance. Ideally, these factors should not exceed 15%. Based on Table 3, the unexplained variance in the 1st contrasts has a value of 2.8713 or 8.8%. This indicates that there is an additional factor not explained by the Rasch model that is quite significant. The unexplained variance in the 2nd contrasts has a value of 2.2901 or 7.0%, indicating the presence of a second factor not explained by the model. These values indicate variance that cannot be explained by the instrument but still contributes to the data variance.

These results indicate that the Rasch model adequately explains the variance in the data, with over 50% of the variance explained by the measures (persons and items) (Andrich and Marais, 2019). However, there is some variance not explained by the model (a total of 50.02%), with some contrasts indicating the presence of additional factors affecting the data. This confirms that although the Rasch model is a good fit, there is still room to improve the model or consider additional factors in the analysis (Andrich and Marais, 2019; Boone and Staver, 2020).

b. Data Fit with the Model

The analysed data is considered to fit the Rasch model if the calculated values of infit MNSQ and outfit MNSQ meet the acceptance criteria, which are within the range of 0.5 to 1.5. For infit ZSTD and outfit ZSTD values, the acceptance criteria are within the range of -2.00 to 2.00. The analysis results for these variables are shown in Table 2.

Table 2. Person and Item Fit with the Rasch Model

	Infit		Outfit	
	MNSQ	ZSTD	MNSQ	ZSTD
Person	0.97	-0.09	1.02	0.00
Item	1.04	-0.11	1.06	-0.04

Based on Table 2, the person infit MNSQ and outfit MNSQ values are 0.97 and 1.02, respectively, very close to 1.00. This indicates that the student response data fits the Rasch model. Similarly, the person infit ZSTD and outfit ZSTD values are -0.09 and 0.00, respectively, indicating very small standard deviations, even zero. This means there are no significant discrepancies between the data and the model. For items, the infit MNSQ and outfit MNSQ values are 1.04 and 1.06, respectively, indicating that the item data fits the Rasch model. The item infit ZSTD and outfit ZSTD values are -0.11 and -0.04, respectively, showing small standard deviations, meaning there are no significant discrepancies between the data and the model. Thus, the data shows that both students (persons) and items have a good fit with the Rasch model from both infit and outfit perspectives. MNSQ values close to 1.00 and ZSTD values close to 0 indicate that this data fits the Rasch model very well.

Table 2 provides information related to infit (inlier-sensitive fit), outfit (outlier-sensitive fit), MNSQ (mean square), and ZSTD (standardized Z-score). Infit indicates sensitivity to misfit affecting respondent ability or item difficulty in a balanced manner. Outfit explains sensitivity to misfit influenced by outliers or extreme values. MNSQ shows the mean square of the deviation, expected to be close to 1.00. ZSTD explains the standard deviation value of the deviation, expected to be close to 0 (Anselmi et al., 2019; Boone and Staver, 2020).

2. Validitas Ditingkat Item

Item-level validity is assessed based on how good, Differential Item Functioning (DIF), and how much and how precise.

a. How Good

How good refers to the content validity of the item, which explains the empirical understanding of respondents towards the item. How good is determined from the results of the item misfit order analysis as shown in Table 3.

Table 3. Item statistics: misfit order

Item	Outfit MNSQ	Outfit ZSTD	PTMEASURE-AL CORR.
Q13	2.55	3.46	0.16
Q12	1.92	2.95	0.55
Q3	1.30	0.83	0.27
Q6	1.17	0.55	0.44
Q16	1.24	-0.67	0.40
Q7	1.12	0.51	0.44
Q5	1.09	0.44	0.20
Q15	0.99	0.05	0.54
Q4	0.94	-0.19	0.69
Q17	0.78	-0.76	0.60
Q1	0.77	-0.84	0.39
Q10	0.74	-0.91	0.44
Q14	0.71	-1.09	0.41
Q11	0.59	-1.57	0.48
Q2	0.52	-1.89	0.55
Q8	0.48	-2.92	0.54

To assess item content validity, several calculated values are considered, such as the outfit MNSQ value, the outfit ZSTD value, and the point measure correlation value. The criteria for each variable are as follows: outfit MNSQ value between 0.5 and 1.5, outfit ZSTD value between -2.0 and +2.0, and point measure correlation value between 0.4 and 0.85 (Andrich and Marais, 2019; Boone and Staver, 2020; Planinic et al., 2019). Based on these criteria, as shown in Table 5, the items that do not meet the outfit MNSQ criteria are Q8, Q12, and Q13. The items that do not meet the outfit ZSTD acceptance criteria are Q8, Q12, and Q13. For the point measure correlation, the items that do not meet the criteria are Q1, Q3, Q5, and Q13. Ideally, an item should meet the established criteria for outfit MNSQ, outfit ZSTD, and point measure correlation values. However, if an item does not meet one of these three criteria, it can still be taken and used as an instrument, provided it meets two out of the three established variable value criteria. For instance, Q1, Q3, and Q5 can be used as instruments because they fall within the acceptance values for outfit MNSQ and outfit ZSTD.

b. Differential Item Functioning (DIF)

DIF in Rasch model analysis refers to the differential functioning of an item across different groups within the tested population. Simply put, DIF occurs when an item in a test or questionnaire has different difficulty levels for different subgroups (e.g., based on gender, race, or ethnic group). In this analysis, an illustration of DIF is obtained as shown in Figure 1.

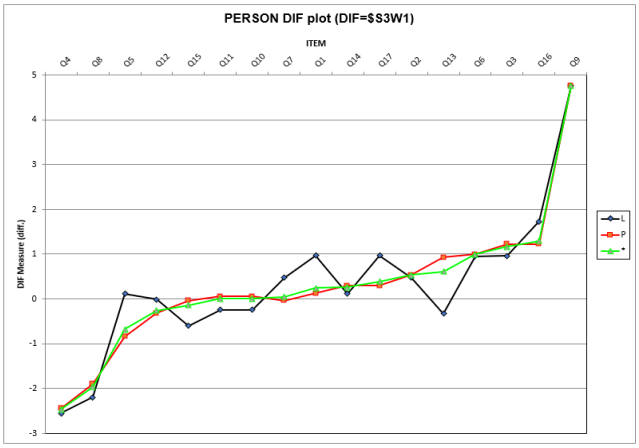

Figure 1. *Differential item functioning*

Figure 1. shows the differences in DIF Measure for each item along the horizontal axis. The three lines represent different groups being compared: the black line with diamonds represents males, the red line with squares represents females, and the green line with triangles represents the combined group of males and females. Based on Figure 1, it is evident that items Q4, Q8, Q10, Q11, Q13, and Q15 show negative DIF values for the male group compared to the female and combined groups. This means that these items are more difficult for males compared to females. Conversely, items Q1, Q5, Q7, Q12, Q16, and Q17 show higher positive DIF values for the male group compared to the female and combined groups, indicating that these items are easier for males. Items with small DIF differences between groups demonstrate good consistency and are not biased towards any group (Bond and Fox, 2015; Rouquette et al., 2019). On the other hand, items with large DIF differences between groups indicate potential bias, as seen with items Q1, Q5, Q7, and Q13. Therefore, these items should be further examined or modified to ensure fairness in testing or potentially not used as instruments if possible.

c. How Much and How Precise

How much refers to the item measure value indicating the difficulty level of the item, while how precise refers to the S.E. value indicating the precision of the item. The calculations for how much and how precise are shown in Table 4.

Table 4. Item measure order

Item	Measure	Model S.E.
Q9	4.76	1.83
Q16	1.29	0.36
Q3	1.17	0.35
Q6	0.99	0.33
Q13	0.60	0.31
Q2	0.53	0.27
Q17	0.39	0.26
Q14	0.26	0.25
Q1	0.24	0.25
Q7	0.05	0.24
Q10	0.00	0.27
Q11	0.00	0.27
Q15	-0.15	0.24
Q12	-0.27	0.23
Q5	-0.68	0.21
Q8	-1.96	0.17
Q4	-2.47	0.20
Mean	**0.28**	**0.36**
P.SD	**1.47**	**0.37**

The calculation of how much is determined by the measure (see Table 4), with the acceptance criteria being values within the range of -2SD to +2SD. Based on Table 4, there are three outlier items (low measurement levels) because their item measure values are outside the acceptance criteria, being either greater than 2SD or smaller than -2SD (Maryati et al., 2019; Wei et al., 2020). These items are Q4, Q8, and Q9, with item Q4 being the easiest to answer and item Q9 being the most difficult to answer. Regarding how precise (the precision level) an item is in measuring; this is indicated by the model S.E. value. An item is considered very precise in measuring if the S.E. value is less than 0.5 and not precise if the S.E. value is greater than 1 (Ling Lee et al., 2021; Sumintono and Widhiarso, 2015). From Table 4, all items have S.E. values within the acceptance range except for Q9, meaning that most items have a high level of measurement precision.

3. Reliabilitas

Pada analisis model Rasch, nilai reliabilitas tidak hanya terdiri atas akumulasi satu nilai. Tetapi menyajikan tiga nilai sekaligus yaitu reliabilitas person, reliabilitas item, dan Alpha Cronbach. Hal ini dapat memberikan penafsiran yang lebih kompleks terhadap suatu instrumen. summary statistik hasil analisis model Rasch dapat dilihat pada Tabel 5.

Table 5. Summary of Rasch Model Analysis Results

	Mean	SD	Separation	Reliability	Alpha Cronbach
Person	-1.84	0.71	1.27	0.62	0.86
Item	0.00	0.98	3.15	0.91	

Table 5. shows the measurement results for the test items. The average item measure is 0.00, which confirms the average item difficulty measured in logits. Since the average item difficulty is 0.00, this means that the item values have been calibrated to center around zero. The standard deviation (SD) of the items is 0.98, which confirms the spread of item difficulties. The higher the SD value, the greater the variation in item difficulties. The separation value is 3.15, which confirms how well the item difficulties can be distinguished (Anselmi et al., 2019; Purnami et al., 2021). A value of 3.15 indicates that this distinction is very good. The item reliability value is 0.91, which confirms the consistency of measuring item difficulties. A value of 0.91 indicates excellent reliability, meaning the measurement of item difficulties is very consistent. On the other hand, the Cronbach's alpha value is 0.86, which is a measure of the internal reliability of the overall test. A value of 0.86 indicates that the test has very good reliability. In general, values above 0.70 are considered good, and values above 0.80 are considered very good (Andrich and Marais, 2019; Sumintono and Widhiarso, 2015).

Table 5 also shows the average person measure of -1.84, which confirms the average ability of respondents measured in logits. The negative value indicates that the average ability of respondents is below the average item difficulty (Boone and Staver, 2020; Laliyo et al., 2022). The standard deviation (SD) of the person measure is 0.27, which confirms the spread of respondent abilities. The higher the SD value, the greater the variation in respondent abilities (Planinic et al., 2019). The separation value is 1.27, which confirms how well the respondent abilities can be separated into different groups. A value greater than 1 indicates that there is a significant difference between higher and lower respondent abilities. The person reliability value is 0.62, which indicates the consistency of measuring respondent abilities. This value is considered weak (Ling Lee et al., 2021; Widhiarso and Sumintono, 2016). Overall, Table 5 shows that the average respondent ability is below the average item difficulty, there is considerable variation in both respondent abilities and item difficulties, the test has good capability in distinguishing respondent abilities and item difficulties, the measurement reliability is excellent for items but weak for respondents, and the internal reliability of the overall test is very good.

In this study, the Rasch model enabled the identification of misfitting items and provided useful information for instrument revision. The analysis results presented by the Rasch model are very comprehensive, as seen from the complex characteristics of the test results for each component. These findings are consistent with previous research that used the Rasch model to evaluate measurement instruments in the field of education. Bond and Fox (2015), and Boone and Staver (2020) demonstrated that the Rasch model is effective in assessing the validity and reliability of instruments, which is also supported by the research of Juandi et al. (2023) in the context of critical and reflective thinking skills. Additionally, research by Laliyo et al. (2022) emphasized the importance of individual item analysis to understand how each item contributes to the overall construct of the instrument. The results of this study show that the Rasch model can identify biased items and improve the internal consistency of the instrument, ensuring that the instrument can be used more effectively to measure reflective thinking skills.

The importance of these findings lies in their contribution to the development of more valid and reliable evaluation instruments in physics education. By using the Rasch model, researchers can ensure that the instruments used genuinely measure the reflective thinking skills necessary for understanding complex physics concepts. The scientific implications of this study provide empirical evidence that the Rasch model is an effective tool for evaluating and improving measurement instruments in the context of higher education. These findings also highlight the importance of continuous revision and development of instruments. By conducting in-depth analysis of item performance and respondent ability, researchers can identify areas needing improvement and ensure that the instruments remain relevant and effective in measurement.

Conclusion

This study successfully analysed the validity and reliability of an instrument for reflective thinking skills on the topic of wave-particle dualism in modern physics lectures using the Rasch model. The analysis results show that the developed instrument has excellent item reliability (0.91) and excellent internal reliability (Cronbach's Alpha 0.86), although respondent reliability remains weak (0.62). This instrument also meets validity criteria, both at the instrument level and the item level. The use of the Rasch model allows for the identification of biased items and improves the internal consistency of the instrument, ensuring that it can be used more effectively to measure students' reflective thinking skills. These findings make a significant contribution to the development of more accurate and reliable evaluation tools in physics education and provide deeper insights into how the instrument functions in different contexts.

Acknowledgements

I would like to express my sincere gratitude to the Indonesia Endowment Fund for Education (LPDP), the Ministry of Finance of the Republic of Indonesia, and Hamzanwadi University for their financial support during my studies at the Indonesia University of Education, Bandung.

References

Alves, E. G., and Santos, A. L. M. (2021). Photoelectric effect: development of a quantitative experiment. *Revista Brasileira de Ensino de Fisica*, *43*, 1–9. https://doi.org/10.1590/1806-9126-RBEF-2021-0146

Andrich, D., and Marais, I. (2019). A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences. In *Springer*. https://doi.org/10.1007/978-981-13-7496-8

Anselmi, P., Colledani, D., and Robusto, E. (2019). A Comparison of Classical and Modern Measures of Internal Consistency. *Frontiers in Psychology*, *10*(December), 1–12. https://doi.org/10.3389/fpsyg.2019.02714

Arabatzis, T. (2017). How Physica Became Physics. In *Science & Education*. Science & Education. https://doi.org/https://doi.org/10.1007/s11191-017-9946-7

Bond, T. ., and Fox, C. (2015). *Applying the Rasch model fundamental measurement in the human sciences (3rd ed.)*. Routledge.

Boone, W. J., and Staver, J. R. (2020). Correction to: Advances in Rasch Analyses in the Human Sciences. In *Springer*. https://doi.org/10.1007/978-3-030-43420-5_21

Hadzhikolev, E., Hadzhikoleva, S., Yotov, K., and Orozova, D. (2020). Models for Multicomponent Fuzzy Evaluation, with a Focus on the Assessment of Higher-Order Thinking Skills. *TEM Journal*, *9*(4), 1656–1662. https://doi.org/10.18421/TEM94-43

Handayani, Y., Rahmawati, R., and ... (2023). Using Rasch Model to Analyze Reliability and Validity of Concept Mastery Test on Electricity and Magnetism Topic. *JIPF (Jurnal Ilmu …*, *8*(2), 226–239. https://doi.org/http://dx.doi.org/10.26737/jipf.v8i2.3877

Juandi, T., Kaniawati, I., Samsudin, A., and Riza, L. S. (2023). Implementing the rasch model to assess the level of students' critical and reflective thinking skills on the photoelectric effect. *Momentum: Physics Education Journal*, *7*(2). https://doi.org/https://doi.org/10.21067/mpej.v7i2.8252

Kanim, S. (2020). Demographics of physics education research. *Physical Review Physics Education Research*, *16*(2), 20106. https://doi.org/10.1103/PhysRevPhysEducRes.16.020106

Laliyo, L. A. R., La Kilo, A., Paputungan, M., Kunusa, W. R., Dama, L., and Panigoro, C. (2022). Rasch Modelling To Evaluate Reasoning Difficulties, Changes of Responses, and Item Misconception Pattern of Hydrolysis. *Journal of Baltic Science Education*, *21*(5), 817–835. https://doi.org/10.33225/jbse/22.21.817

Ling Lee, W., Chinna, K., and Sumintono, B. (2021). Psychometrics assessment of HeartQoL questionnaire: A Rasch analysis. *European Journal of Preventive Cardiology*, *28*(12), E1–E5. https://doi.org/10.1177/2047487320902322

Maryani, I., Prasetyo, Z. K., Wilujeng, I., Purwanti, S., and Fitrianawati, M. (2021). HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers. *Journal of Turkish Science Education*, *18*(4), 674–690. https://doi.org/10.36681/tused.2021.97

Maryati, M., Prasetyo, K. un Z., Wilujeng, I., and Sumintoni, B. (2019). Measuring Teachers' Pedagogical Content Knowledge Using Many-Facet Rasch Model. *Cakrawala Pendidikan*, *38*(3), 452–464. https://doi.org/10.21831/cp.v38i3.26598

Nitriani, N., Darsikin, D., and Saehana, S. (2022). Kolb's learning style analysis in solving HOTS questions for prospective physics teacher students. *Momentum: Physics Education Journal*, *6*(1), 59–72. https://doi.org/10.21067/mpej.v6i1.5593

Planinic, M., Boone, W. J., Susac, A., and Ivanjek, L. (2019). Rasch analysis in physics education research : Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 20111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Purnami, W., Ashadi, Suranto, Sarwanto, Sumintono, B., and Wahyu, Y. (2021). Investigation of person ability and item fit instruments of eco critical thinking skills in basic science concept materials for elementary pre-service teachers. *Jurnal Pendidikan IPA Indonesia*, *10*(1), 127–137. https://doi.org/10.15294/jpii.v10i1.25239

Rouquette, A., Hardouin, J., Vanhaesebrouck, A., Se´bille, V., and Coste, J. l. (2019). Differential Item Functioning (DIF) in composite health measurement scale : Recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. *Plos One*, 1–16. https://doi.org/10.1371/journal. pone.0215073

Salido, A., and Dasari, D. (2019). The analysis of students' reflective thinking ability viewed by students' mathematical ability at senior high school. *Journal of Physics: Conference Series*, *1157*(2). https://doi.org/10.1088/1742-

6596/1157/2/022121

Samsudin, A., Rusdiana, D., Efendi, R., Fratiwi, N. J., Aminudin, A. H., and Adimayuda, R. (2021). Development of Predict-Observe-Explain (POE) Strategy Assisted by Rebuttal Texts on Newton's Law Material with Rasch Analysis. *Tadris: Jurnal Keguruan Dan Ilmu Tarbiyah*, *6*(1), 103–115. https://doi.org/10.24042/tadris.v6i1.7641

Setiawan, J., Sudrajat, A., Aman, and Kumalasari, D. (2021). Development of higher order thinking skill assessment instruments in learning Indonesian history. *International Journal of Evaluation and Research in Education*, *10*(2), 545–552. https://doi.org/10.11591/ijere.v10i2.20796

Sumintono, B., and Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Asesmen Pendidikan*. Bandung: Trim Komunikata.

Sun, Y., and Latora, V. (2020). The evolution of knowledge within and across fields in modern physics. *Scientific Reports*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-68774-w

Syahidi, K., Jufri, A. W., Doyan, A., Rokhmat, J., and Sukarso, A. A. (2023). Penguatan Literasi Sains dan Pendidikan Karakter pada Pembelajaran IPA Abad 21. *Kappa Journal*, *7*(3), 538–542. https://doi.org/https://doi.org/10.29408/kpj.v7i3.25036

Wei, S., Chee, C., Looi, K., and Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students : a Rasch model measurement analysis. *Journal of Computers in Education*, *0123456789*. https://doi.org/10.1007/s40692-020-00177-2

Widhiarso, W., and Sumintono, B. (2016). Examining response aberrance as a cause of outliers in statistical analysis. *PAID*, *98*, 11–15. https://doi.org/10.1016/j.paid.2016.03.099