Journal of Educational Research and Evaluation
Volume 8, Issue 3, 2024, pp. 443-451
P-ISSN: 2597-422x E-ISSN: 2549-2675
Open Access: https://doi.org/10.23887/jere.v8i3.83688

Application of the Rasch Model to the Evaluation of Biological Problems

Ida Elfira[1], Abdul Razak[2*], Syamsurizal[3], Hendra[4]

[1,2,3,4] Department of Biology, Padang State University, Padang, Indonesia

ARTICLE INFO

Article history:
Received June 29, 2024
Accepted August 10, 2024
Available online August 25, 2024

Kata Kunci :
Analisis pertanyaan, Validitas, Reliabilitas, Pengukuran Rasch

Keywords:
Question analysis, Validity, reliability, Rasch measurement

ABSTRAK

Evaluasi kemampuan siswa dalam mata pelajaran Biologi sering kali menghadapi tantangan dalam memastikan bahwa instrumen yang digunakan benar-benar mengukur apa yang seharusnya diukur. Analisis data yang digunakan adalah model rasch, yang merupakan bagian dari teori respons butir, menawarkan pendekatan yang lebih akurat dalam menganalisis data evaluasi. Model ini memungkinkan identifikasi item yang kurang sesuai atau terlalu sulit, serta memberikan informasi yang lebih mendalam tentang kemampuan individu siswa. Melalui analisis instrumen dan item soal dari 30 soal pilihan ganda yang menekankan pada kemampuan berpikir kritis, tujuan penelitian ini adalah untuk menilai kualitas soal Biologi kelas X SMA. Pengujian dilakukan dengan software WinStep dan analisis pemodelan Rasch pada 166 siswa dari lima SMA/SMK. Metode penelitian yang digunakan adalah kuantitatif. Hasil investigasi instrumen menunjukkan bahwa isu tersebut memiliki ketergantungan yang sangat baik dan unidimensi instrumen ditunjukkan oleh perubahan yang tidak dapat dijelaskan pada diferensiasi pertama. Hasil tersebut menunjukkan bahwa tes Rasch dapat membantu pendidik dalam memilih soal yang memenuhi kriteria evaluasi dan berkontribusi terhadap peningkatan mutu pendidikan nasional. Implikasi penelitian ini adalah Dengan menerapkan Model Rasch, pendidik dapat mengembangkan alat evaluasi yang lebih akurat dan adil, yang mampu mengidentifikasi kesulitan dan kelebihan siswa secara individual.

ABSTRACT

Evaluation of student abilities in Biology subjects often faces challenges in ensuring that the instruments used actually measure what they are supposed to measure. The data analysis used is the Rasch model, which is part of item response theory, offering a more accurate approach in analyzing evaluation data. This model allows the identification of items that are inappropriate or too difficult, as well as providing more in-depth information about individual student abilities. Through analysis of instruments and question items from 30 multiple choice questions that emphasize critical thinking skills, the aim of this research is to assess the quality of Biology questions for class X SMA. Testing was carried out using WinStep software and Rasch modeling analysis on 166 students from five high schools/vocational schools. The research method used is quantitative. The results of the instrument investigation show that the issue has very good dependability and the unidimensionality of the instrument is indicated by unexplained changes in the first differentiation. These results show that the Rasch test can help educators in selecting questions that meet the evaluation criteria and contribute to improving the quality of national education. The implications of this research are By applying the Rasch Model, educators can develop more accurate and fair evaluation tools, which are able to identify individual student difficulties and strengths.

1. INTRODUCTION

The curriculum does not alone determine learning; assessment methods also play a role (Ibarra-Sáiz et al., 2021; Nurmatova & Altun, 2020). The assessment process in education is an integral part of the learning process. The quality of assessment is closely related to assessment, because assessment questions require the interpretation and application of criteria (Schellekens et al., 2023; Setiabudi et al., 2019) Assessment findings are also considered a reliable source to ensure that programs and curricula are of high quality and aligned with the goals and mission of learning institutions. It will be difficult to determine whether learning progress has been achieved or not without evaluation (Elawed, 2024; Wahidin & Romli, 2019). Teachers can use more detailed assessment results to find out how well a student performed compared to other students who took the same test; shows how students' abilities develop over time in specific knowledge and skills; show evidence of understanding of specific subject matter, knowledge, or ideas; and predicting student performance in the future (Wahidin & Romli, 2019; Wahyuni, 2019). By using Bloom's cognitive domain as a model, one approach to formulating questions

*Corresponding author.
E-mail addresses: idaelfira5567@gmail.com (Ida Elfira)

in this scenario is to use quality learning evaluation to support the growth of students' critical thinking abilities (Cecilio-Fernandes et al., 2019; Sadhu et al., 2020).. Higher order thinking questions including the cognitive domains of analysis (C4), synthesis (C5), and judgment (C6) have been designed in relation to the ecological material. Bloom's Taxonomy is a logically structured framework that outlines the cognitive abilities students need to understand information thoroughly and meaningfully. Applying Bloom's taxonomy in the classroom increases students' understanding of critical thinking and analytical skills (Afriana et al., 2018; Ali, 2021)

The questions developed also contain critical thinking criteria based on the views of Robert Ennis who identified five indicators of critical thinking skills: providing basic explanations (basic clarification), constructing basic skills (basic support), drawing conclusions, offering additional explanations (advanced clarification), and develop strategies and tactics (strategy and tactics). In addition, these questions build basic skills, summarize, provide additional explanations, and manage strategies and techniques (Cecilio-Fernandes et al., 2019; Kapas, 2021). One of the most crucial abilities in the twenty-first century is the ability to think critically (Daskalopoulou, 2020; Elawed, 2024). This will help someone select and analyze information in everyday life (Wahidin & Romli, 2020). Therefore, students must be educated in this competency, one of which is through the questions asked of them. Critical reasoning, one of the components of the Pancasila student profile in the autonomous curriculum, is aligned with critical thinking skills (Indonesian Ministry of Education and Culture, Research and Technology, 2021) Announced as a solution to the intense global competition for human resources in the context of Society 5.0.

One of the stages in an effort to help children develop higher order thinking skills (HOTS) is by asking questions that include critical thinking components. Higher order thinking skills (HOTS) are cognitive talents that require a more complex reasoning process than simply memorizing. Critical, imaginative, analytical and introspective thinking are part of HOTS (Farisi et al., 2021; A. Fitriani et al., 2020). The aim of High Order Thinking Skills (HOTS) is to improve children's thinking abilities in higher categories related to being receptive to information, creative in solving problems, and being able to make decisions in the right circumstances (Munar et al., 2022). Problems that arise in the HOTS type can be found in Bloom's taxonomy at levels C4, C5, and C6 (S. S. Fitriani et al., 2019; Handayani & Rahmawati, 2020). Teachers rarely, if ever, study test questions. As a result, they are not aware of the quality of the tests given to students. This should motivate teachers to focus more on the validity and reliability of the test questions they will be given to their students, because this is an appropriate way to assess student learning. The requirements for validity and reliability of the instrument must be met so that the test questions can be trusted and suitable for use (Hauer et al., 2020; Ibarra-Sáiz et al., 2021). The idea of validity is that teachers must use valid tests. A valid test is a test that can only measure what it is designed to measure. Dependability is the quality of being dependable and dependable. A test is said to be reliable if it consistently provides the same results under the same conditions (Karim et al., 2022; Mokshein et al., 2018). Previous research findings stated that dependability refers to the consistency of the scores obtained. Rasch measurement as a measurement model and data analysis method Walter R. Borg (1993) (Bond, 2015). One other metric that can be applied to determine the level of work on test question instruments. The Rasch measurement model was chosen as an alternative creation tool for educational exams, helping educators in evaluating students and improving the caliber of their questions through the analysis carried out. Validity, reliability and analysis of test results will be taken into consideration in assessing the quality of the questions. All analyzes can be completed quickly with Rasch modeling and the results are useful and high quality measurement data (Munar et al., 2020; Nurmatova & Altun, 2020). The aim of this research is to apply the Rasch Model in the evaluation of Biology problems in order to increase the validity and reliability of assessment instruments. This research aims to identify and analyze the level of difficulty and quality of the question items used in the Biology test, as well as to understand how students' abilities are distributed in relation to these questions. By applying the Rasch Model, this research also aims to provide recommendations that can be used by educators in developing assessment instruments that are more accurate, fair and appropriate to students' abilities. In addition, it is hoped that this research can contribute to the development of more objective evaluation methods, which can increase the effectiveness of the learning process in the field of Biology.

2. METHODS

This research method uses quantitative research methods. This research focused on analyzing the test instrument in the form of multiple choice questions for Biology class The questions developed include learning objectives related to ecosystem concepts and interactions in ecosystems, food webs or food

chains as well as energy flows and pyramids in ecosystems. This research involved 166 respondents from 5 high schools/vocational schools in the city of Padang, West Sumatra. Data analysis using the Rasch model with the help of Winstep 3.73 software will be used to analyze the instrument and analyze the questions by looking at the level of difficulty of the size of the questions, the level of suitability of the questions and detection of the suitability of the questions. biased question items, then the data will be interpreted to see the feasibility of the questions that have been prepared. After analysis, appropriate questions can be used by teachers as alternative questions to assess students' understanding of ecosystem material in class X SMA/SMK. The research method used in applying the Rasch Model to evaluating Biology problems includes several main stages, namely data collected through tests or questionnaires containing a series of Biology questions. Respondents usually consist of students who are studying Biology subjects. Second, assessment instruments are developed and prepared in accordance with the applicable Biology curriculum. The questions are created based on competency standards that must be achieved by students, taking into account variations in levels of difficulty. Third, the data that has been collected is analyzed using the Rasch Model. Special software such as Winsteps or RUMM2030 can be used to perform this analysis. The Rasch model will evaluate the level of difficulty of each question item, as well as the student's ability to answer the questions. Fourth, the instrument used will be tested for validity (whether it measures what it is supposed to measure) and reliability (whether the measurement results are consistent). The Rasch model provides detailed information regarding validity and reliability through fit statistics, item characteristic curves, and person-item maps. Fifth, the results of the analysis are then interpreted to understand the extent to which the instruments used meet assessment quality standards. These findings are then reported in the form of research which includes recommendations for improving the assessment instrument in the future.

3. RESULT AND DISCUSSION

Results

Summary of instrument statistics

At the instrument level, the results of Rasch measurement analysis can be used to determine the reliability and validity of the instrument. The consequences of this instrument inspection will provide extensive data and will provide data to emissions producers to make appropriate, reasonable and logical choices in the context of a thorough and top to bottom investigation. The consequences of instrument-level examination are presented in Table 1.

Table 1. Summary of Instrument Statistics

Summary of Measured Person			
Infit		**Outfit**	
MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
0.98	0.0	1.07	0.1
Separation	2.30	Person Reliability	0.84
Person raw score-to-measure correlation			1.00
Cronbach alpha (KR-20) person raw score "Test" Reliability			0.86
Summary of Measured Item			
Infit		**Outfit**	
MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
1.00	-0.2	1.07	-0.3
Separation	5.12	Person Reliability	0.97

The validity of the test items will also be tested to determine the extent to which they can assess what is actually being measured in accordance with the predetermined conceptual understanding or definition. The validity value can be used to determine whether the question is valid or not. Validity testing in Rasch analysis will use principal component analysis of standardized residual variance (in Eigenvalue units) presented in Table 2. For item analysis, item size detection was carried out using standard deviation (SD) values combined with the average logit presented in Table 3.

Table 2. Standardized Residual Variance Table (In Eigenvalue Units)

	Eigenvalue	**Empirical**	**Modeled**
Total raw variance in observations	44.1	100.0%	100.0%
Raw variance explained by measures	14.1	31.9%	31.2%

	Eigenvalue	Empirical	Modeled
Raw variance explained by persons	5.9	13.4%	13.1%
Raw variance explained by items	8.2	18.5%	18.1%
Raw unexplained variance (total)	30.0	68.1%	68.8%
Unexplained variance in 1st contrast	3.4	7.7%	11.3%

Table 3. Difficulty Level of Question Details

Standard deviation value (SD)		**1.03**
Question number	**Measure**	**Question Group**
5, 20, 28, 10, 15.	2.17, 2.07, 1.98, 1.26, 1.12.	Very difficult
9, 26, 27, 18, 14, 30, 7, 6.	0.92, 0.92, 0.88, 0.82, 0.53, 0.29, 0.26, 0.19.	Difficult
1, 3, 25, 16, 17, 29, 19, 23, 4, 21, 8, 11, 13, 24.	-0.05, -0.47, -0.47, -0.50, -0.50, -0.60, -0.73, -0.77,-0.80, -0.80, -0.87, -0.91, -0.94, -0.98.	Easy
12, 22, 2.	-1.25, -1.29, -1.49.	Very easy

The purpose of item fit analysis is to evaluate how well the question items fit the model. The ability of a query item to collect measurements regularly or not is explained by the suitability of the item. If it is found that the question item is inappropriate, this indicates that students may have a wrong perception about it. Teachers can utilize this information to improve their teaching and prevent misunderstandings when they teach the material again. Parameters used to determine the degree of suitability of an item. Outfit mean-square value, outfit z-standard, and point gauge correlation value. It can be determined that the question item is of poor quality and must be changed or replaced if the three requirements of the question item are not met. The following standards can be applied to determine whether a non-matching question item is appropriate: Acceptable MNSQ clothing value: 0.5<MNSQ<1.5. Acceptable values for clothing ZSTD are -2.0<ZSTD<+2.0 and 0.4<Pt Measure Corr<0.85. If one of the question items is found where the MNSQ and Pt Measure Corr values do not meet the criteria but the ZSTD value meets the criteria, then the item is still considered fit, meaning the item is still maintained. In this test question, the discrepancies are presented in Table 4.

Table 4. Item Mismatch

Question number	**Outfit**		**Pt-Measure Corr**
	MNSQ	**ZSTD**	
10	2.97	8.0	-0.29
28	2.69	4.7	-0.14
20	1.81	2.6	-0.03
5	1.61	-	0.24
18	-	2.9	0.15
15	-	-	0.25
30	-	2.2	0.37
14	-	-	0.37
3	-	-	0.40
9	-	-	0.40
27	-	-	0.38
24	-	-2.2	-
29	-	-2.6	-
11	-	-2.1	-
17	-	-2.5	-
23	-	-3.0	-
16	-	-3.6	-

The same structures should be measured in assessments conducted in other languages. However, certain test takers may benefit more from certain item features than others, such as item errors and content. Item bias is the term for this. One indicator of valid measurement is the absence of bias in the instrument and question items. If one individual attribute is more helpful to an instrument or question item than someone with another quality, then the instrument or question item is considered biased. In this biology question instrument, prejudice is measured using only one variable, namely the gender category (male and female). If it is known that the probability value of a question item is less than 5% (0.05), then it is said to be biased. To guarantee that the test questions are unbiased and able to show

equality of build for each test taker, a DIF study is needed. This research is usually conducted while analyzing measurement systems, adapting tests for various cultural contexts, and validating test results in general. Therefore, it can be concluded that DIF inquiry is a standard procedure in the process of verifying measurement findings scores. A graph of question items that have a gender bias is presented in Figure 1.

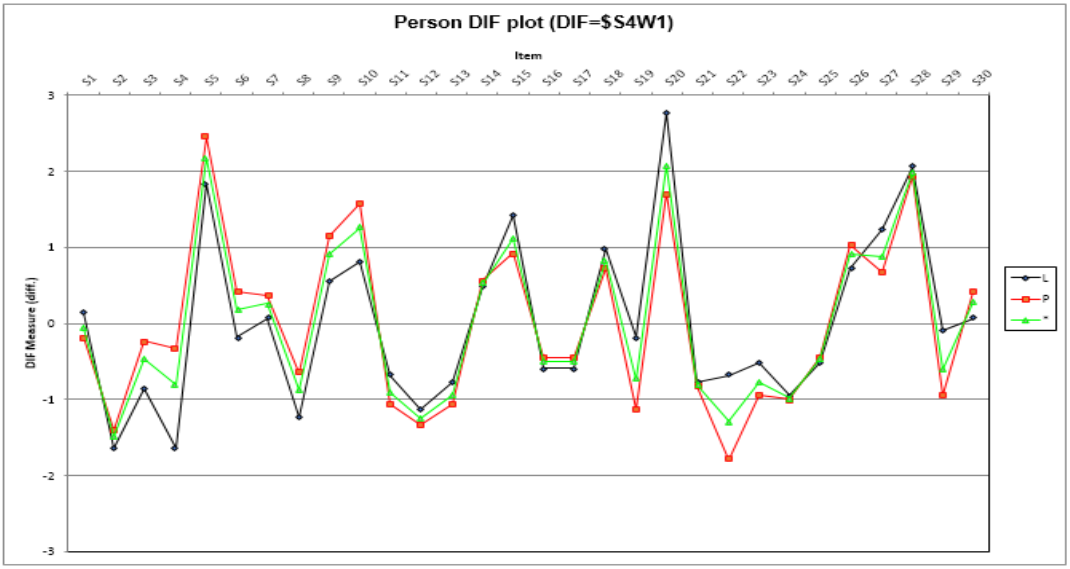

Figure 1. Human Biology DIF Question Instrument

Discussion

Table 1 displays an overview of the instrument statistics. indicates that a person's measurement value is 1. The logit value of 0.0 is smaller than the average value. This shows the tendency of students' talents to exceed the level of complexity of the questions. Based on the above criteria, the figure is 0.8, the Cronbach's alpha value which measures reliability as an interaction between individuals and question items as a whole with a value of 0.86 indicates very good reliability (Rezai et al., 2020; Rukli & Ma'rup, 2020). This shows that the internal consistency of the science question instrument is generally very good. For the individual dependency test, the consistency of substitute answers is 0.84, indicating the consistency of substitute answers is "very good". The object dependency test score of 0.97 indicates that the nature of inquiry is in the "extraordinary" class. This shows that the question instruments developed generally have very high reliability coefficient values. Various information that can be described is INFIT MNSQ information, OUTFIT MNSQ, INFIT ZSTD and OUTFIT ZSTD information. For individual tables, the typical gains from INFIT MNSQ and OUTFIT MNSQ are 0.98 and 1.07 separately and the gains from INFIT ZSTD and OUTFIT ZSTD are 0.00 and 0.10. Since all the scores are close to 1.00, the nature of the questions seems good. This is according to the setting where the best value is 1.00 (Sadhu et al., 2020; Schellekens et al., 2023), so the quality will be better if the value is closer to 1.00. Similarly, the item tables for INFIT MNSQ and OUTFIT MNSQ have values of 1.00 and 1.07, respectively. Additionally, the item tables for INFIT ZSTD and OUTFIT ZSTD have values of -0.2 and -0.3. Next, a value of 2.30 is displayed for the separation of people included in the adequate category, and a value of 5.12 for the separation of items included in the special category. This separation value is used to measure how effectively the test instrument can differentiate respondents with different skills. The quality of the instrument will increase with a higher separation value because it is able to differentiate groups of respondents and groups of items (Setiabudi et al., 2019; Steinke & Fitch, 2024).

The Rasch analysis validity test uses standardized principal component analysis of residual variance (in Eigenvalue units), with a minimum score of 20% considered fulfilled, a minimum score of 40% as very good, and a minimum score of 60% for extraordinary requirements. To determine whether the instrument can measure the uniformity of a single dimension with an adequate level of confidence, validity will be assessed using the RASCH model to ensure all items belong to the same dimension (Wahidin & Romli, 2019; Wahyuni, 2019). The magnitude of the variation in the measured data is shown by the raw variance value. The test questions met the validity requirements and were considered suitable for use, based on the Rasch test findings which showed that the variance/raw eigenvalue of the instrument was 31.2%. The findings show that the unidimensionality of the instrument with a minimum value of 20% is met. In contrast, the value of unexplained variance is 11.3%, which is less than 15%. This shows that the question instrument is unidimensional, meaning it can measure something that should be able to be measured. In Table 2. In addition, it is clear that the observed raw variance explained by person

(RVEP) and observed raw variance explained by item (RVEI) values almost match the model values empirically. This shows that the Rasch model is a good fit for the data.

Item size

Starting with the level of difficulty of the question item (measurement item), the standard deviation (SD) and average logit values are used to carry out item analysis. For example, 0.0 logit +1SD represents a challenging class of problems, 1.0 logit -1SD represents easy problems, and less than -1SD represents very easy problems. Question Response Theory states that one of the characteristics of test questions is their level of difficulty. One of the characteristics of test questions that are scaled using ability level parameters is the level of difficulty (Rezai et al., 2020; Walker, 2019). The standard deviation of this question instrument is 1.03, meaning that the very difficult question category is above 1.03, the difficult question category is between 0.0 to 1.03, the easy question category is between 0.0 to (-1.03), and the questions in the very easy category are less than -1.03. In Table 3 above, the question items are grouped based on their level of difficulty. Educators can measure students' understanding of subject matter by using the Rasch model to determine the level of difficulty of test items. This will assist teachers in finding ways to help students understand the difficulties of assessments that are considered challenging, especially teaching engineering approaches. Rasch analysis will be of great benefit to educators in their learning process activities.

In the science question instrument which is broken down using the Rasch model, it can be seen that the very challenging inquiry classes are distributed in questions number 5 (C5/reasoning point of view that determines recognizing or planning possible response standards), number 20 (C4/reasoning that determines the perspective that makes each angle is unique), number 28 (reasoning point of view that determines dissecting the source of information), number 10 (reasoning point of view that determines differentiating or describing the size of a conceivable response), and number 15 (C5/technique of point of view definition decisive reasoning). Then, for the difficult question category, question number 9 (Aspect C4/critical thinking identifying or formulating criteria for possible answers), question number 26 (Aspect C4/critical thinking identifying or formulating criteria for possible answers), question number 27 (C4/critical thinking aspect thinking aspect identifying or formulating possible answer criteria), question number 18 (C4/critical thinking aspect identifying or formulating questions), question number 14 (C5/critical thinking aspect considering or deciding), question number 30 (C4/From the circulation of investigative information which is really challenging and troublesome considering the Rasch test information, it is generally seen that students actually need strengthening in the perspective of decisive reasoning, especially the ability to differentiate or determine possible response models, find differences in every corner, dissect sources of information, consider sources of information choice, considering and choosing activities. This will be input for teachers to focus more on the decisive reasoning angle to ensure students develop their reasoning cycle, especially critical thinking.

The appropriate thing makes sense regardless of whether the object of inquiry is capable of regularly taking estimates. Based on Table 4. As can be seen, questions number 10, 28, and 20 are three questions that do not meet the fit item requirements. These three questions do not meet the rules of outfit value implying squares, z-standard outfit and point size connections so they tend to be considered to be less good investigations so they have to be repaired or replaced. Meanwhile, for question number 5, the MNSQ and Pt Measure Corr scores do not meet the standard but the ZSTD score meets the model, so question number 5 is still considered fit so question number 5 can still be maintained. with. For questions number 18 and 30 it tends to be less suitable for the reason that it does not meet the Outfit Zstd and Pt Measure Corr requirements, however for the Outfit Mnsq standard these two items are still included as far as possible so the things referred to in numbers 18 and 30 are still maintained.. Meanwhile, the other 11 inquiry items are specific inquiry numbers 15, 14, 3, 9, 27, 24, 29, 11, 17, 23, 16 because they do not meet the model in one of the ZSTD values or ZSTD values. Pt Measure Corr appreciates that the things asked can be maintained and do not need to be changed so that they can still be used (Afandi et al., 2018; Tivani & Paidi, 2016). The instruments used for biology questions. Regarding the questions, there are 8 questions that have a probability value below 5% (0.05), namely questions number 1, 10, 18, 20, 21, 26, 28 and 29 so these questions are classified as one-sided questions. These biased questions indicate that one gender benefits from the question, so the question needs to be changed to prevent harm to that particular gender group. Apart from that, it can be seen from graph 2 that there are six questions—questions 1, 18, 20, 21, 28, and 29—that are easier for female students to do than male students (the female curve is below the male curve). Meanwhile, it is easier for male students to do questions number 10 and number 26 than female students (men's bends are below women's bends), this could be a concern for question makers to limit or try to turn off assessment questions. which has a one-sided nature so as not to harm one association (Sekolah et al., 2021; Sumartini, 2022). Examining potential trends when surveying student work can help uncover concerns among teachers evaluating how viewpoints that are generally underrepresented in schools can be overlooked This may have an impact on academic careers and may impact overall ranking. Therefore, to avoid bias and create an ideal assessment system, practical recommendations are needed in education World. Teachers are obliged to take action in an effort to avoid bias (Febriana Sulistya Pratiwi., 2022; Hizhwati et al., 2022; Nawawi & Wijayanti, 2018). To reduce or eliminate bias in assessment, it is important to employ a number of key strategies, such as using a variety

of test formats that suit different students' learning preferences and test materials that take into account different aspects of a person's identity, such as gender, encouraging students verbally and providing feedback. in writing. The research findings on the application of the Rasch Model to the evaluation of Biology problems show a significant increase in the accuracy and fairness of assessment compared to previous research that used traditional approaches, such as classical analysis (CTT). In previous research, assessment instruments often show limitations in measuring students' abilities consistently and fairly, especially in terms of item discrimination and sensitivity to differences in students' abilities. In contrast, the findings from this research reveal that the Rasch Model is able to provide more in-depth and detailed analysis, such as the ability to identify items that do not function well (misfitting items) and provide a clear person-item map. Additionally, the Rasch Model allows for a more accurate examination of how item difficulty is proportional to student ability, which was not found in previous research. Thus, this research provides a significant new contribution to the literature, showing that the application of the Rasch Model can produce evaluation instruments that are more valid, reliable, and responsive to variations in student abilities, compared to previous assessment methods. The implications of this research cover various important aspects in education, especially in improving the quality of assessment and learning instruments. By applying the Rasch Model, educators can develop more accurate and fair evaluation tools, which are able to identify individual student difficulties and strengths. This allows for more appropriate adaptation of teaching materials and methods according to students' learning needs. Apart from that, another implication is the creation of a more transparent and accountable evaluation, because the Rasch Model allows a more detailed analysis of student performance and the quality of the question items. Ultimately, this research can encourage the development of educational policies that focus more on measuring student competency objectively and consistently, as well as encouraging increased educational standards in the field of Biology.

One of the limitations of this research is the need for a fairly large and representative sample to obtain accurate and generalizable analysis results. Using the Rasch Model also requires a deep technical understanding and access to specialized software, which may be a challenge for educators who do not have a strong background in statistics or technology. In addition, the application of the Rasch Model is more effective on tests with homogeneous item formats; therefore, a wide variety of items may require additional adjustments in the analysis. As a recommendation, further research should involve larger and more diverse samples to ensure results that are more robust and can be applied more widely. In addition, training for educators in the use of the Rasch Model and its software needs to be improved so that they can implement this approach independently. Further research is also recommended to explore the integration of the Rasch Model with more diverse assessment formats, such as essay questions or practical assignments, to understand the potential and limitations of this model in more complex assessment contexts. Thus, future research may provide more comprehensive guidance for the development of more adaptive and effective evaluation instruments in the field of Biology.

4. CONCLUSION

To ensure that the quality of the questions created meets the requirements of an assessment tool, educators who carry out assessments are strongly advised to use Rasch measurement as one type of measurement. The validity and reliability of the instrument, which can be seen from the general analysis of the instrument, can be a consideration in evaluating questions. Test item analysis includes determining the level of difficulty of the item, suitability of the item, and detection of bias. One way for educators to encourage the creation of a quality curriculum through a long-term national education assessment system is through evaluation tests with measurable quality, with the aim of making the nation's children smarter.

5. REFERENCES

Afandi, A., Junanto, T., & Afria, R. (2018). Implementasi digital-age literacy dalam pendidikan abad 21 di Indonesia. *Prosiding SNPS (Seminar Nasional Pendidikan Sains)*, *8*(2), 113–120.

Afriana, N., Halim, A., & Syukri, M. (2018). Analisis Karakteristik Keterampilan Berpikir Kritis Siswa dalam Menyelesaikan Soal Ujian Nasional. *Jurnal Penelitian Pendidikan IPA*, *7*(2), 196. https://doi.org/10.29303/jppipa.v7i2.627.

Ali, R. (2021). Differential item functioning pada tes multidimensi. In *Buku Pegangan Teori Respon Item* (Vol. 3, pp. 67–86). https://doi.org/10.4324/9780203115190-13.

Cecilio-Fernandes, D., Bremers, A., Collares, C. F., Nieuwland, W., Vleuten, C., & Tio, R. A. (2019). Menyelidiki kemungkinan penyebab bias dalam tes kemajuan terjemahan: Pedang bermata satu. *Jurnal Pendidikan Kedokteran Korea*, *31*(3), 193–204. https://doi.org/10.3946/kjme.2019.130.

Daskalopoulou, A. (2020). Memahami dampak evaluasi siswa yang bias: analisis titik-temu pengalaman akademisi dalam konteks pendidikan tinggi di Inggris. *Studi Di Pendidikan Tinggi*, *2*(2), 1–12.

https://doi.org/10.1080/03075079.2024.2306364.
Elawed, A. F. E. (2024). Dampak Penerapan Pengukuran Dan Evaluasi Terhadap Peningkatan Mutu Keluaran Pendidikan Universitas. *Administrasi Pendidikan: Teori Dan Praktek*, *30*(4), 6415–6429. https://doi.org/10.53555/kuey.v30i4.2399.
Farisi, A., Hamid, A., & Melvina. (2021). Pengaruh Model Pembelajaran Problem Based Learning terhadap Kemampuan Berpikir Kritis dalam Meningkatkan Hasil Belajar Siswa pada Konsep Suhu dan Kalor. *Jurnal Ilmiah Mahasiswa (JIM) Pendidikan Fisika*, *2*(3), 283–287.
Febriana Sulistya Pratiwi. (2022). Biologi. *5*, ארץ(8.5.2017), 2005–2003. https://dataindonesia.id/sektor-riil/detail/angka-konsumsi-ikan-ri-naik-jadi-5648-kgkapita-pada-2022.
Fitriani, A., Zubaidah, S., & Hidayati, N. (2020). Kualitas berpikir kritis siswa: Survei sekolah menengah atas di Bengkulu, Indonesia. *JPBI (Jurnal Pendidikan Biologi Indonesia*, *8*(2), 142–149. https://doi.org/10.22219/jpbi.v8i2.18129.
Fitriani, S. S., Yusuf, Y. Q., & Zumara, A. (2019). Penggunaan ranah kognitif dalam pertanyaan: Persepsi mahasiswa dan dosen perguruan tinggi negeri di Aceh. *Jurnal Studi Bahasa Dan Linguistik*, *17*(1), 122–138. https://doi.org/10.17263/jlls.903359.
Handayani, Y., & Rahmawati, R. (2020). Penggunaan Model Rasch untuk Menganalisis Reliabilitas dan Validitas Uji Penguasaan Konsep pada Topik Listrik dan Magnet. *JIPF (Jurnal Ilmu*, *8*(2), 226–239. https://doi.org/https://journal.stkipsingkawang.ac.id/index.php/JIPF/article/view/3877.
Hauer, K. E., Park, Y. S., Bullock, J. L., & Tekian, A. (2020). Penilaian Saya Bias!" Pengukuran dan Pendekatan Sosial Budaya untuk Mewujudkan Keadilan Penilaian dalam Pendidikan Kedokteran. *Kedokteran Akademik*, *98*(8), 16– 27. https://doi.org/10.1097/ACM.0000000000005245.
Hizhwati, D., Susilo, S., Amirullah, G., & Supardi, S. (2022). Persepsi Guru Biologi terhadap Mobile Learning dalam Pembelajaran Biologi. *EduBiologia: Biological Science and Education Journal*, *2*(2), 106. https://doi.org/10.30998/edubiologia.v2i2.12973.
Ibarra-Sáiz, M. S., Rodríguez-Gómez, G., & Boud, D. (2021). Kualitas tugas penilaian sebagai penentu pembelajaran. *Asesmen Dan Evaluasi Di Perguruan Tinggi*, *46*(6), 943–955. https://doi.org/10.1080/02602938.2020.1828268.
Kapas, K. (2021). Komunitas Belajar Kecil Baru: Temuan dari Literatur Terkini. *Ed*, *459*(12), 539.
Karim, S. A., Sudiro, S., & Sakinah, S. (2022). Memanfaatkan analisis soal tes untuk mengetahui tingkat kesukaran dan daya pembeda suatu tes buatan guru. *EduLite: Jurnal Pendidikan, Sastra Dan Budaya Inggris*, *6*(2), 256. https://doi.org/10.30659/e.6.2.256-269.
Mokshein, S. E., Ishak, H., & Ahmad, H. (2018). Penggunaan model pengukuran rasch dalam pengujian bahasa Inggris. *Cakrawala Pendidikan*, *38*(1), 16–32. https://doi.org/10.21831/cp.v38i1.22750.
Munar, A., Winarti, W., Nai'mah, N., Rezieka, D. G., & Aulia, A. (2020). Meningkatkan Keterampilan Berpikir Tingkat Tinggi (Hots) pada Anak Usia Dini dengan Menggunakan Buku Cerita Bergambar. *AL-ISHLAH: Jurnal Pendidikan*, *14*(3), 4611–4618. https://doi.org/10.35445/alishlah.v14i3.2224.
Nawawi, S., & Wijayanti, T. F. (2018). Pengembangan asesmen biologi berbasis keterampilan berpikir kritis terintegrasi nilai Islam. *Jurnal Inovasi Pendidikan IPA*, *4*(2), 136–148. https://doi.org/10.21831/jipi.v4i2.21265.
Nurmatova, S., & Altun, M. (2020). Tinjauan Komprehensif Integrasi Taksonomi Bloom untuk Meningkatkan Dampak Pedagogis Pendidik EFL Pemula. *Jurnal Bahasa Inggris Dunia Arab*, *14*(3), 380–388. https://doi.org/10.24093/awej/vol14no3.24.
Rezai, A., Namaziandost, E., Miri, M., & Kumar, T. (2020). Bias demografis dan keadilan penilaian di kelas: wawasan dari para dosen universitas Iran. *Pengujian Bahasa Di Asia*, *12*(1), 145. https://doi.org/10.1186/s40468-022-00157-6.
Rukli, R., & Ma'rup, M. (2020). Dampak Siswa Menilai Tingkat Kesukaran Soal Tes Terhadap Minat dan Motivasi Belajar Matematika. *Jurnal Sains Pendidikan Dan Linguistik Internasional Randwick*, *3*(2), 229–240. https://doi.org/10.47175/rielsj.v3i2.476.
Sadhu, S., Ad'hiya, E., & Laksono, E. W. (2020). Menjelajahi dan membandingkan validitas isi dan asumsi teori modern penilaian terpadu: Studi literasi berpikir kritis-kimia. *Jurnal Pendidikan IPA Indonesia*, *8*(4), 570–581. https://doi.org/10.15294/jpii.v8i4.20967.
Schellekens, L. H., Kremer, W. D. J., Schaaf, M. F., Vleuten, C. P. M., & Bok, H. G. J. (2023). Antara teori dan praktik: persepsi pendidik terhadap kriteria kualitas penilaian dan dampaknya terhadap pembelajaran siswa. *Perbatasan Dalam Pendidikan*, *8*(Juni), 1–9. https://doi.org/10.3389/feduc.2023.1147213.
Sekolah, T., Atas, M., & Di, S. M. A. (2021). *Siswa Pada Tingkat Sekolah Menengah Atas (Sma) Di*. *23*(23l), 13.
Setiabudi, A., Mulyadi, & Puspita, H. (2019). Analisis Validitas dan Reliabilitas Tes Buatan Guru (Studi Kasus pada Kelas XI SMA N 6 Bengkulu. *Jurnal Pendidikan Dan Pengajaran Bahasa Inggris)JEET*,

3(4), 522–532.

Steinke, P., & Fitch, P. (2024). Meminimalkan Bias Saat Menilai Hasil Karya Siswa. *Penelitian & Praktek Dalam Penilaian*, *12*(2), 87–95.

Sumartini, A. (2022). Penerapan Taksonomi Digital Bloom Pada Masa Belajar Di Rumah Oleh Guru SMK Di Kalimantan Barat. *Jurnal Pendidikan Indonesia*, *3*(08), 748–760. https://doi.org/10.59141/japendi.v3i08.998.

Tivani, I., & Paidi, P. (2016). Pengembangan LKS biologi berbasis masalah untuk meningkatkan kemampuan pemecahan masalah dan karakter peduli lingkungan. *Jurnal Inovasi Pendidikan IPA*, *2*(1), 35. https://doi.org/10.21831/jipi.v2i1.8804.

Wahidin, D., & Romli, L. A. M. (2019). Perkembangan Berpikir Kritis Siswa pada Kompetisi Sains dan Matematika Nasional di Indonesia: Kajian Deskriptif. *Jurnal Pendidikan IPA Indonesia*, *9*(1), 106–115. https://doi.org/10.15294/jpii.v9i1.22240.

Wahyuni, A. R. (2019). Pengaruh Model Pembelajaran REACT (Relating, Experiencing, Applying, Cooperating and Transferring) terhadap Keterampilan Berpikir Kritis dan Keterampilan Kreativitas Siswa Milenial di SMA. *Jurnal Internasional Ilmu Sosial Dan Penelitian Manusia*, *04*(12), 3954–3958. https://doi.org/10.47191/ijsshr/v4-i12-69.

Walker, C. M. (2019). Apa DIFnya? mengapa analisis fungsi item diferensial merupakan bagian penting dari pengembangan dan validasi instrumen. *Jurnal Penilaian Psikoedukasi*, *29*(4), 364–376. https://doi.org/10.1177/0734282911406666.