



Uncovering Students' Problem-Solving Gaps in Basic Chemistry Through Rasch Analysis

Sri Rejeki Dwi Astuti^{1*}, Pramudya Dwi Aristya Putra², Nur Ahmad³, Firdha Yusmar⁴,
Jadnika Dwi Rakhmawan Amrullah⁵, Ika Nur Fitriani⁶

¹⁻⁵Department of Science Education, Faculty of Education, Universitas Jember, Indonesia

⁶School of Chemistry, The University of Sydney, Australia

*Correspondence to: srirejeki.fkip@unej.ac.id

Abstract: Problem-solving skills are fundamental competencies in science education, particularly in basic chemistry, as they enable students to implement conceptual knowledge to real-life phenomena. This study aims to examine the psychometric properties of the problem-based chemistry test and to describe students' problem-solving skills. The test consisted of five essay-based problem-solving items administered to 123 first-year biology education students enrolled in a general chemistry course. Data were analyzed using EFA with SPSS 25 and Rasch Model with Winsteps Software version 3.73. The results indicated that all items fit the Rasch Model and demonstrated good item reliability, with a range of difficulty levels. However, person reliability was low, suggesting that the test had limited sensitivity in differentiating students' ability levels. Analysis of students' abilities revealed that most students were classified as having low to very low problem-solving skills, while a smaller proportion demonstrated moderate to high ability. A substantial number of persons misfit were identified, indicating irregular response patterns that may be influenced by test-taking behavior or the open-ended nature of the items. These findings suggest that the test is suitable for formative and diagnostic assessment of problem-solving skills in basic chemistry; however, further refinement and additional items are suggested to improve measurement precision.

Keywords: assessment; basic chemistry; problem-solving skills; rasch analysis

Recommended citation: Astuti, S. R. D., Putra, P. D. A., Ahmad, N., Yusmar, F., Amrullah, J. D. R., & Fitriani, I. N. (2026). Uncovering Students' Problem-Solving Gaps in Basic Chemistry Through Rasch Analysis. *Journal of Innovation in Educational and Cultural Research*, 7(3), 564-578.

INTRODUCTION

Higher education plays a crucial role in advancing the Sustainable Development Goals (SDGs) by equipping students with the knowledge and skills needed to address global challenges. Universities and colleges increasingly align their curricula with sustainability principles, fostering interdisciplinary learning and critical thinking (Abo-Khalil, 2024). Higher education institutions are well-positioned to prepare students who can contribute to achieving these global objectives as future members of society. Thus, embedding education for sustainable development (ESD) in curricula is crucial for equipping graduates (Price et al., 2024; Rodríguez-Zurita et al., 2024). By implementing active learning, universities can engage students through sustainable development initiatives (Nansen, 2024; Rosak-Szyrocka & Knop, 2024). Active learning can be implemented through cooperative models, such as problem- or project-based learning, which use real-world problems to introduce the course topic. It can enhance knowledge and sustainable practices (Biancardi, et al., 2023).

To fulfill the demands of the SDGs, universities and colleges must emphasize 21st-century skills such as creativity, critical thinking, problem-solving skills, collaboration, communication, self-regulation, and socio-emotional skills (Alturki & Aldraiweesh, 2023; Stehle & Peters-Burton, 2019; Tan et al., 2017; Thornhill-Miller et al., 2023). It prepares students to engage in global challenges, apply academic knowledge in real-life contexts, and contribute to sustainable development. To prepare students for the challenges of the 21st century, educational content must be integrated into the analysis, synthesis, and evaluation steps of the learning process. This ensures that students develop their higher-order thinking skills and can participate in informed scientific discussions. By doing so, students can better understand how scientific arguments are constructed and develop the proficiency needed to engage with scientific content (Küçükaydın & Ayaz, 2025).

Critical thinking is a fundamental skill for understanding the complexities of sustainable development, as it involves analyzing real-world issues and devising innovative solutions (Martin et al., 2020). With critical thinking skills, students can develop effective arguments to identify real-world problems related to SDG issues and formulate solutions to address these issues. Critical thinking not only enhances problem-solving skills but also serves as a mediator in developing these skills by enhancing students' confidence (Boran & Karakus, 2022; Tasgin & Dilek, 2023). Enhancing critical thinking significantly improves problem-solving skills because it requires identifying problems, evaluating relevant information, and applying knowledge to solve them (Amanda et al., 2024; Doleck et al., 2017; Nold, 2017).

Problem-solving skills are the highest level of thinking and involve important skills that encompass complex processes, including reading, data collection, solving mathematical problems, and finding solutions (Demitra & Sarjoko, 2018; Riyadi et al., 2021). It plays a crucial role in the science classroom as a learning objective, fostering competence in addressing new challenges and proficiency in solving practical problems. Additionally, it serves as an instructional tool to transform the science classroom, making it a more active, relevant, and inspiring experience (Yerushalmi & Eylon, 2015). It is in accordance with the current curriculum in Indonesia, Merdeka curriculum (Merdeka Belajar-Kampus Merdeka), which aims to provide schools with the ability to design their own curricula to be more flexible in accordance with local demands and conditions, as well as adjust to SDGs, preparing students to face the opportunities and challenges in global economy and society (Fatmawati et al., 2024). One of the goals of the current curriculum is to enhance problem-solving skills essential for addressing global challenges. Problem-based and project-based learning are implemented in the classroom to facilitate students' problem-solving skills. Problem-based learning emphasizes students collaborating in groups to solve problems and provide solutions to challenging issues in real-world contexts related to their course subjects, thereby acquiring meaningful knowledge (Asyhari & Sifa'i, 2021; Hmelo-Silver, 2004). Project-based learning emphasizes the collaborative process of investigating a problem, culminating in a performance, presentation, or product as an outcome of learning activities (Artama et al., 2023; Karan & Brown, 2022). Both of these learning models utilize real-world issues related to the course subject as topics, allowing students to learn the materials.

Basic chemistry, a compulsory university course, has a unique role in supporting the SDGs. It aligns with sustainability goals by explaining chemistry-related phenomena in daily life. Thus, teachers can encourage students to connect theoretical concepts with real action by incorporating SDGs related to chemistry classes, stimulating a sense of innovation and responsibility in addressing global issues. Therefore, students must learn by exploring problems and solving them using the principles of chemistry. Integrating green chemistry principles into chemistry classrooms has been explored as a means to promote sustainability. This approach equips students with technical expertise and cultivates their environmental awareness and problem-solving abilities. Helping students develop their problem-solving skills is one of the most challenging aspects of beginning chemistry courses, as analogies are not particularly helpful for specific problems such as atomic structure and chemical bonding (Sa-ngiemjit et al., 2024). However, other topics can relate to daily-life phenomena, such as stoichiometry, reaction rates, chemical equilibrium, chemical solutions, electrochemistry, hydrocarbons, and colloidal systems. Additionally, teachers can implement a learning model that supports problem-solving skills.

Assessment plays a crucial role in the learning process, serving as a guide to navigate student progress and evaluate skill acquisition. In chemistry, assessments should adapt to reflect the increasingly complex skills required of students, including critical thinking and problem-solving abilities needed to navigate real-world phenomena (Underwood et al., 2018; Ying & Tiemann, 2024). Assessing problem-solving skills in basic chemistry is crucial for developing students' competencies, which are necessary for both academic success and real-world applications. It not only enhances understanding of chemical concepts but also fosters critical thinking and creativity, which are essential in the 21st-century educational landscape as they enable students to tackle complex chemical problems effectively (Chi & Ninh, 2018) and prepare them for future challenges in STEM fields (Ying & Tiemann, 2024). Moreover, practical skills such as laboratory experimentation must also be assessed to ensure students are proficient in hands-on techniques (Eidesen et al., 2023; Riaz et al., 2023). In addition, a good assessment in chemistry classrooms should effectively measure students' ability to think critically and solve problems holistically. This can be achieved by designing tests that evaluate not only the recall of theoretical concepts but also the application and analysis of chemical principles to real-world scenarios. Moreover, open-ended questions and case studies can be used to assess students' ability to generate critical thinking and problem-solving skills, as well as communicate complex scientific ideas effectively. By incorporating these assessment strategies, teachers can gain a deeper understanding of students' learning outcomes and provide targeted feedback to support their ongoing development.

Various assessment tools, such as the Collaborative Problem-Solving tool, have been developed to evaluate students' abilities in practical scenarios, ensuring a comprehensive understanding of chemistry (Bueno, 2014; Shadle et al., 2012; Ying & Tiemann, 2024). These tools are essential in promoting higher-order thinking skills and preparing students for real-world scientific challenges. However, there remains a lack of specific assessment tools designed to measure problem-solving skills in basic chemistry, particularly for students majoring in biology education. Basic chemistry is crucial for comprehending biological processes, as it underpins life at the molecular and cellular levels. Thus, it equips biology students with the conceptual tools needed to explain how and why biological systems function, as life processes are chemical reactions that involve the structure and function of biomolecules. Preliminary observations in a basic chemistry course indicated that students had difficulty applying conceptual understanding to problem-solving tasks, particularly when required to interpret data and justify solutions. Initial analysis of students' responses revealed inconsistencies in problem-solving performance that were not adequately captured by existing assessment instruments. Therefore, it is important to develop and validate a problem-solving skills test that aligns with the needs and characteristics of

biology education students, thereby fostering both disciplinary competence and scientific reasoning. Based on these research backgrounds, we have identified crucial matters such as assessing problem-solving skills in the basic chemistry classroom. In addressing this problem scope, this objective result is guided by the following research question: 1) How well do the problem-based basic chemistry test items fit the Rasch measurement model? 2) What are the item difficulty and person ability distributions of students' problem-solving skills based on Rasch analysis? To overcome these drawbacks, this research aims to explore item parameters of a problem-based basic chemistry test and investigate students' problem-solving abilities.

METHODS

The study utilized a quantitative approach (Creswell, 2009) to analyze the item parameters of the instrument and person characteristics using modern analysis (Rasch model) (Davis & Boone, 2021). The study explores item difficulty, item fit, person measure (ability), person fit, and person misfit. The study involved 123 students from the Biology Education Department at Universitas Jember, selected using a purposive sampling technique, with consideration given to students who enrolled in basic chemistry courses during their first year. They came from four different classes taught by the same lecturer. Thus, we assumed that the learning process in all classes was the same. The participants' demographic profile, including Biology Education 1 (N=31, F=90.33%; M=9.67%), Biology Education 2 (N=31, F= 87.10%; M= 12.90%), Biology Education 3 (N=31, F= 96,78%; M=3.22%), and Biology Education 4 (N=30, F=87,10%; M=13.33%)

The instrument utilized in the study was the lecture-based test designed to measure students' problem-solving skills in basic chemistry. It is an essay test comprising five questions, developed based on problem-solving and learning indicators in basic chemistry. Therefore, it can be utilized to measure students' problem-solving skills in basic chemistry. The question grid for problem-solving skills in basic chemistry tests is presented in Table 1.

Table 1. Question grid

Learning Indicators	Problem-Solving Indicators	Item Indicators	Item number
Solving problems related to the reaction rate	Understanding the problem	Provided a case of cancer therapy, students should analyze the cancer therapy based on the reaction of the inhibition rate of medicine for inhibiting the metastatic cancer	1
	Checking the solution	In the case of industrial ammonia formation, students should analyze the rate of ammonia formation by comparing rate laws to increase the reaction rate.	2
Solving problems related to chemical equilibrium	Planning a strategy	Provided a case study of the influence of hyperventilation on blood pH, students should explain the effect of the equilibrium shift in the blood and offer a solution to prevent an increase in blood pH.	3
Solving problems related to chemical solutions	Understanding the problem	Provided a case of the presence of alcohol levels in the blood can damage the body. Thus, the students should determine the blood alcohol level in the human body and explain how alcohol affects the body.	4
Solving the problem related to stoichiometry	Checking the solution	In the case of an explosive reaction with nitroglycerin, students should determine the level of the product compound in the reaction and provide a solution for storing nitroglycerin so it does not explode quickly.	5

The Rasch model was employed to analyze students' abilities, especially problem-solving skills. Several assumptions must be met before conducting a Rasch model analysis, including unidimensionality and local independence. An Exploratory Factor Analysis was performed to test the unidimensionality assumption of the Rasch model in SPSS version 25. After the unidimensional assumption was fulfilled, the data analysis continued using the Rasch model, with Winstep version 3.73 employed to determine test characteristics and students' abilities. It was conducted to verify validity and reliability using unidimensionality and local independence tests, based on person and item reliability criteria. Furthermore, the Wright map illustrated the relationship between test item difficulty and students' ability. Moreover, the analysis results using the Rasch model include item parameters (item fit, item difficulty, and test information function), person parameters (person fit, person misfit, and person measure/person ability), and item bias detection.

RESULT AND DISCUSSION

Assessing a student's ability should use a valid and reliable instrument with appropriate item parameters, particularly item difficulty. A valid instrument can perform the measurement function since it can measure the purpose. Meanwhile, reliable instruments can yield similar measurements when tested on different samples. Moreover, good instruments should have a balanced distribution of item difficulty, with easy (30%), moderate (40%), and complex items (30%) (Fauzie et al., 2021; Sudjana, 2017). After the instrument is proven valid and reliable and has a suitable item parameter, it can accurately measure the student's ability. However, analyzing the item and the students' abilities separately takes a long time. Therefore, it can be conducted simultaneously with item parameter analysis using the Rasch Model.

Before conducting the Rasch Model analysis, it is essential to test the model's assumptions, including unidimensionality and local independence. Unidimensionality is the assumption that the instrument only measures one factor or variable and that there is no correlation between items in this instrument. (Ziegler & Hagemann, 2015). Exploratory factor analysis (EFA) is a feasible approach for testing the unidimensionality hypothesis (Hattie, 1985; Ziegler & Hagemann, 2015). EFA started with the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO-MSA) analysis. It is a crucial tool in assessing the suitability of data for factor analysis (Lee et al., 2023; Nkansah, 2018; Costales et al., 2022). The primary function of KMO-MSA was to assess sample size adequacy, while Bartlett's test was used to determine the presence of correlation among variables. It helps determine the dataset's suitability for this statistical technique by considering factors such as data dimensionality, variable influence across dimensions, and variable correlation structure (Nkansah, 2018). However, KMO-MSA values can be influenced by factors such as the number of response categories, sample size, and the types of correlation analysis (Lee et al., 2023).

The KMO-MSA value is 0.631. KMO-MSA value within the range of 0.6 to 0.7 is generally considered mediocre sampling. However, it was desired for factor analysis (Nkansah, 2018; Watkins, 2018). Moreover, Approx. Chi-Square and the significance of Bartlett's test of sphericity were 70.249 (df=10), and sig=0.000, which is less than 0.05, indicating a significant correlation among the variables in the data (Sigudla & Maritz, 2023). After verifying the sample size was adequate and confirming it was sufficient for factor analysis, the analysis proceeded with an exploratory approach using principal component analysis (PCA) to determine the number of factors via eigenvalue interpretation. Factor analysis is used to investigate the correlation between variables by utilizing correlation tests. Hence, new variables called factors are produced.

Table 2. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of variance	Cumulative %
1	1.930	38.608	38.608
2	1.015	20.296	58.904
3	0.949	18.976	77.880
4	0.636	12.726	90.606
5	0.470	9.394	100.00

As stated in the Kaiser criterion, the number of factors is equal to the number of components with eigenvalues higher than one, since it can explain more information than a single variable, whereas an eigenvalue less than 1 contains less information (Beavers et al., 2013; Jamshidi et al., 2019; Patyka et al., 2021). Table 2 presents that only two components, with eigenvalues higher than 1, account for a cumulative 58.904% of the total variance. The unidimensional assumption is difficult to fulfill in practice. However, it was considered fulfilled if a dominant factor was used to measure students' ability. The existing dominant factor is indicated by the cumulative percentage of the first factor of more than 20% (Hambleton et al., 1991; Reckase, 1979; Smits et al., 2011; Wu et al., 2013).

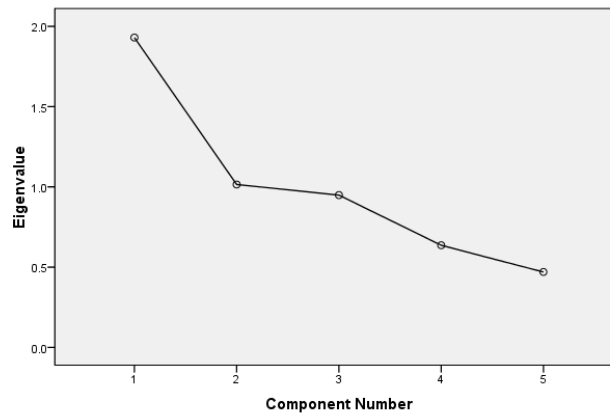


Figure 1. Scree Plot of Exploratory Factor Analysis

The result of the factor analysis is visualized in a scree plot to illustrate the eigenvalue pattern presented in Figure 1. The scree plot is a standard tool in exploratory factor analysis and principal component analysis (PCA) for visualizing the relative importance of factors/components based on eigenvalues, which indicate the variance accounted for by each factor. In Figure 1, it was observed that the factors after the second dot were less than one. Hence, it was concluded that the test has two factors. The first and second factors explained 38.608% and 20.296% of the total variance. Thus, the total variance explained was 58.904 %. The factors were named based on the literature, items, and expert views. According to the findings, of these two factors, the dominant factor measured by the test is problem-solving skills in basic chemistry, as the test is composed of a question grid based on problem-solving skills indicators and learning indicators in basic chemistry. Moreover, another factor contributing to the test response is the test's administrative factors, including students' anxiety and motivation during the test (Balogun et al., 2017; Bolbolian et al., 2021; Dikmen, 2023; Fr'echette-Simard et al., 2023). Thus, the test explains 58.904% of students' problem-solving skills. Therefore, the unidimensional assumption has been fulfilled. After the unidimensional assumption was met, the analysis proceeded with a Rasch model analysis to determine item characteristics and students' problem-solving skills. Before analyzing students' problem-solving skills, we determined item fit and item difficulty, as students' abilities can be assessed once the item characteristics have been identified. The results of the statistical summary for item and person (sample) analysis, as presented by Winstep, are shown in Table 3.

Table 3. Statistic Summary of Measured Item and Person

		Item Measured	Person Measured
Total Score	Mean	31.4	1.3
	S.D	26.6	1.1
Count	Mean	123.0	5.0
	S.D	0.0	0.0
Real RMSE Measure		0.35	2.02
	Mean	0.00	-2.01
Model Error	S.D	2.01	2.09
	Mean	0.35	1.89
True SD	S.D	0.03	0.71
		1.98	0.53
Infit MNSQ	Mean	0.98	0.89
	S.D	0.19	1.47
Infit ZSTD	Mean	-0.1	-0.1
	S.D	1.1	0.9
Outfit MNSQ	Mean	1.16	1.14
	S.D	0.46	2.49
Outfit ZSTD	Mean	0.1	-0.1
	S.D	1.1	1.0
Reliability		0.97	0.07
Separation		5.63	0.26

S.E. of Person MEAN = 0.19

Person RAW Score-to-Measure Correlation= 0.98

Cronbach Alpha (KR-20) Person RAW Score "Test" Reliability = 0.56

As indicated in Table 3, the reliability of items in the problem-solving test is 0.97 ($\alpha > 0.99$) with a separation index of 5.88 (separation > 3.00), which means that the problem-solving test has excellent internal consistency and can discover a wide measurement range (several-level measurement) (Abdellatif, 2023; Linacre, 2004; Post et al., 2022; Tavakol & Dennick, 2011; Yasin et al., 2023). The separation index complements reliability by quantifying the number of distinct ability strata a test can identify among respondents. A high separation index, typically above 3.0, suggests the ability to distinguish between multiple levels of ability. In contrast, a high-reliability score (close to 1.0) indicates that the model performs consistently across those levels. In educational assessments, these indices are essential for understanding the robustness and effectiveness of an instrument in measuring diverse skill levels, particularly in contexts where multiple skill strata need to be identified (Linacre, 2023). In Rasch, a high-reliability index is unreliable because large sample sizes can yield high-reliability indices. Thus, a separation index is needed to demonstrate the item-level validity and strengthen the test reliability explanation (Bond & Fox, 2015; Souza et al., 2017).

Furthermore, person reliability was analyzed to determine whether students as test respondents have consistent or varied skills. Table 3 shows that students' reliability is 0.07 and their separation index is 0.26. This means the person's reliability was low due to a few skill variations along the logit line. Thus, many students were measured to have the same ability (θ). Moreover, person separation was minimal (separation < 2.0), so the instrument may not be sensitive enough to classify students' abilities into high, medium, and low ability levels (Post, et al., 2022). Person reliability was independent of sample size but dependent on the sample ability variances resulting from a more comprehensive sample ability range, which, in turn, led to a more diverse level of sample ability. Moreover, the length of the test and the number of categories or scoring scales per item also depended on the person's reliability (Linacre, 2023). In addition, it can be anticipated by composing more items to distinguish between high and low ability levels (Souza et al., 2017).

Before analyzing students' abilities, it is essential to analyze item characteristics, such as item fit and item difficulty level. Item fit indicates the extent to which an item conforms to the model, indicating whether it functions properly in measurement (Sumintono & Widhiarso, 2015; Linacre, 2023). Through item fit, the consistency of students' response patterns to a particular item can be evaluated (Centoni & Maruotti, 2021). If the response to an item is inconsistent, it signifies that the item is a misfit (inappropriate) for analysis using the Rasch Model.

Item fit and misfit in test items play a crucial role in the accuracy and reliability of assessments. Detecting misfits is essential as they can stem from various sources, such as special knowledge, cheating, guessing, fatigue, warming up, or faking (Dünya, 2024). While item removal due to misfit is generally undesirable, its practical consequences are often overlooked, with minor effects on person allocation and gender-difference estimations (Fährmann et al., 2022). In the context of item response theory, misfits can manifest as deviations in item discrimination parameters, impacting the precision of parameter estimation and the linking of scales over time (Fischer et al., 2021).

An item is considered to fit the model if it meets the criteria of mean-square (MNSQ) and z-standard value (ZSTD) (Bond & Fox, 2015; Hrnjičić & Alihodžić, 2024). If the item does not meet these criteria, it is deemed a misfit with the analysis model. As a general rule, an MNSQ value greater than 2.0 indicates a detrimental effect on the measurement system, while a value between 1.5 and 2.0 suggests a lack of meaningfulness in the measurement. MNSQ values between 0.5 and 1.5 are considered beneficial for measurement, and an MNSQ value below 0.5, though not useful for measurement, does not disrupt the measurement system (Linacre, 2022; Pitaloka et al., 2023). In addition to the MNSQ value, the ZSTD and Pt mean correlation values can also serve as criteria for item fit. An item is considered to fit if the ZSTD value falls within the range of -2 to +2 and the Pt mean correlation score falls within the range of 0.4 to 0.85 (Boone et al., 2014; Pitaloka et al., 2023). The result of item fit analysis were item 1 (Outfit MNSQ= 1.44; Outfit ZSTD= 0.90; PT-Mean Corr= 0.44), item 2 (Outfit MNSQ= 0.99; Outfit ZSTD= 0.10; PT-Mean Corr= 0.54), item 3 (Outfit MNSQ= 0.52; Outfit ZSTD= -2.00; PT-Mean Corr= 0.71), item 4 (Outfit MNSQ= 0.99; Outfit ZSTD= 0.10; PT-Mean Corr= 0.57), item 5 (Outfit MNSQ= 1.89; Outfit ZSTD= 1.30; PT-Mean Corr= 0.64). The Outfit-MNSQ values for items 1, 2, 3, and 4 in the problem-solving test fall within the range of $0.5 \leq \text{MNSQ} \leq 1.5$, while item number 5 exceeds 1.5. The ZSTD and Pt mean corr values are also above the criteria. If only two of these three criteria are met, the item may still be considered a fit because it has a slightly high outfit MNSQ but falls within the acceptable range for other statistics; it might still be deemed acceptable, depending on the specific circumstances and the overall fit of the test. (Müller, 2020). Therefore, all items fit the Rasch Model analysis. Items that fit the model can serve as effective data-collection instruments for specific purposes. Consequently, these items can effectively measure students' problem-solving abilities in the basic chemistry course.

In measurement instruments, misfit items must be monitored because they do not significantly contribute to test score reliability (Fährmann et al., 2022; Tesio et al., 2024; Zubairi & Kassim, 2006). The misfit of an item with the model can indicate a misconception among test participants about the item, as revealed by abnormal response patterns; thus, the test is not unidimensional (Effatpanah et al., 2024; Kusuma et al., 2022;

Soeharto & Csapo, 2022; Sumintono & Widhiarso, 2015). A misfit item is less effective in measurement. Therefore, misfit items can be removed (if there are many items in the test) or improved (if there are few items in the test) to prevent further misconceptions. After all items were confirmed to fit the Rasch model, the analysis proceeded with determining the difficulty level of each item. This step was not undertaken during the item development phase but was analyzed after the instrument was administered based on students' responses to each test item (Abdellatif, 2023; Boateng et al., 2018). Items that most students answered correctly were considered easy, whereas those that only a few students answered correctly were considered difficult (Rezigalla, et al., 2024). The Rasch model was used to analyze the difficulty levels of the questions per established standards. The item difficulty level parameter (b) represents the estimated difficulty level of an item, expressed in logits (logarithm of the odds unit). An item is considered acceptable if its difficulty level falls within the range of -2.0 logits to $+2.0$ logits (Baker, 2001; Hambleton & Swaminathan, 1985; Planinic et al., 2019). Items with a difficulty level below -2.0 logits are categorized as very easy, while those with a difficulty level above $+2.0$ logits are categorized as very difficult. The difficulty level of an item affects students' ability by influencing the probability of a correct response. Items with a great difficulty level require a higher level of ability to be answered correctly, whereas items with a low difficulty level require a lower level of ability.

In this research, items are categorized as easy if the b value is less than -1 logit, indicating that both low- and high-ability students have a high probability of responding correctly. Items are categorized as medium if their difficulty level falls within the interval $-0.5 < b < +0.5$. Conversely, items are considered problematic if the b value exceeds $+1$, suggesting that low-ability students are unlikely to respond correctly (Adedoyin & Mokobi, 2013). The estimated item difficulty levels were 1.58, 0.87, 0.28, 1.20, and -3.92 logits for Items 1 through 5, respectively. Based on these values, Items 1, 2, and 3 were classified as difficult, Item 4 as moderate, and Item 5 as very easy. Item number 5 has a b value less than -2.00 logits, making it unsuitable due to its very easy difficulty level. However, the other items exhibit an appropriate level of difficulty.

Furthermore, an analysis of students' problem-solving skills was conducted to assess their ability to respond to items in basic chemistry tests. Ability (θ) was obtained based on the student's response to the test. The classification of ability is almost the same as the classification of item difficulty level because if a student has a value of θ that is the same as b , then that student will be able to respond correctly to that item or the item below it (Planinic et al., 2019). Therefore, student ability can be classified as a very high ability if $\theta > +3.00$, high ability if $+3.00 < \theta < +1$, moderate ability if θ is in the range $-1.00 < \theta < +1.00$, low ability if $-3.00 < \theta < -1.00$, and very low ability if $\theta < -3.00$. Students' ability estimates (θ) ranged from -4.82 to $+2.14$ logits. Based on these estimations, 18 students were classified as having high ability ($+1.00 < \theta < +3.00$), 17 students as medium ability ($-1.00 < \theta < +1.00$), 59 students as low ability ($-3.00 < \theta < -1.00$), and 29 students as very low ability ($\theta < -3.00$).

A total of 18 students with abilities in the range $+1.00 < \theta < +3.00$ logits had high abilities. Students with high ability will be able to answer all the questions in the test correctly because all the questions in the instrument have a difficulty level index lower than the student's ability ($b < \theta$). They can solve problems and provide the best solution for the problem offered in the test. Students with an ability $\theta < -3.00$ logits had very low ability and were unable to answer the questions correctly. Because the test is an open-ended essay question, students with low abilities may not be able to respond correctly, resulting in a perfect score on all question items. Students at lower ability levels are disadvantaged when test items are cognitively complex or open-ended, as multiple-choice tests measure only surface information in particular skills and are generally easier than open-ended question tests (Bond & Fox, 2015; Gebremichael et al, 2025; Polat, 2020). Given these constraints, tailoring item difficulty to a range of abilities is essential for fair assessment, as misalignment may lead to biased performance evaluations (Wright & Stone, 1979). By matching item difficulty to the learner's ability level, IRT-based assessments can avoid "ceiling" or "floor" effects when questions are too easy or too hard for some students.

For example, in adaptive testing, items adjust in real time based on the student's previous responses, which enhances the accuracy of ability measurement (Contrino et al., 2024; Ntumi, 2025). The difficulty parameter in IRT reflects the point at which 50% of a population with a particular ability level will correctly answer an item (Yang & Kao, 2014), which allows teachers to measure both how students respond to varying levels of challenge and how their abilities correspond to their performance on items of different levels (Lahza et al., 2022). This alignment enables more effective assessments and supports targeted learning interventions, as students receive tasks that are just challenging enough to improve their skills without overwhelming them, fostering sustained engagement and learning (Lestari & Yusuf, 2025; Rad, 2025). Teachers can effectively address individual differences in ability, contributing to a fairer and more supportive assessment environment.

Table 4. Person Misfit

Students code	θ (logit)	Outfit MNSQ	ZSTD	PT Corr	No	Students code	θ (logit)	Outfit MNSQ	ZSTD	PT Corr
4	-0.14	9.90	2.90	-0.29	35	55	-2.28	0.07	-0.80	0.98
94	-0.14	9.90	3.00	-0.41	36	56	-2.28	0.07	-0.80	0.98
97	-2.29	9.90	3.10	-0.40	37	60	-2.28	0.07	-0.80	0.98
112	-2.29	9.90	3.10	-0.40	38	61	-2.28	0.07	-0.80	0.98
120	-0.14	9.90	3.00	-0.40	39	62	-2.28	0.07	-0.80	0.98
58	-2.29	7.13	2.40	-0.28	40	63	-2.28	0.07	-0.80	0.98
113	-2.29	7.13	2.40	-0.28	41	66	-2.28	0.07	-0.80	0.98
98	-2.29	5.92	2.20	-0.22	42	68	-2.28	0.07	-0.80	0.98
116	-2.29	5.92	2.20	-0.22	43	69	-2.28	0.07	-0.80	0.98
84	-2.29	3.73	1.60	-0.07	44	70	-2.28	0.07	-0.80	0.98
101	2.14	1.51	0.80	0.07	45	71	-2.28	0.07	-0.80	0.98
5	-2.29	0.07	-0.80	0.98	46	73	-2.28	0.07	-0.80	0.98
9	-2.29	0.07	-0.80	0.98	47	75	-2.28	0.07	-0.80	0.98
10	-2.28	0.07	-0.80	0.98	48	76	-2.28	0.07	-0.80	0.98
18	-2.28	0.07	-0.80	0.98	49	77	-2.28	0.07	-0.80	0.98
19	-2.28	0.07	-0.80	0.98	50	78	-2.28	0.07	-0.80	0.98
21	-2.28	0.07	-0.80	0.98	51	79	-2.28	0.07	-0.80	0.98
24	-2.28	0.07	-0.80	0.98	52	80	-2.28	0.07	-0.80	0.98
25	-2.28	0.07	-0.80	0.98	53	81	-2.28	0.07	-0.80	0.98
27	-2.28	0.07	-0.80	0.98	54	82	-2.28	0.07	-0.80	0.98
31	-2.28	0.07	-0.80	0.98	55	83	-2.28	0.07	-0.80	0.98
32	-2.28	0.07	-0.80	0.98	56	85	-2.28	0.07	-0.80	0.98
35	-2.28	0.07	-0.80	0.98	57	86	-2.28	0.07	-0.80	0.98
36	-2.28	0.07	-0.80	0.98	58	89	-2.28	0.07	-0.80	0.98
37	-2.28	0.07	-0.80	0.98	59	103	-2.28	0.07	-0.80	0.98
38	-2.28	0.07	-0.80	0.98	60	114	-2.28	0.07	-0.80	0.98
40	-2.28	0.07	-0.80	0.98	61	115	-2.28	0.07	-0.80	0.98
41	-2.28	0.07	-0.80	0.98	62	117	-2.28	0.07	-0.80	0.98
42	-2.28	0.07	-0.80	0.98	63	123	-2.28	0.07	-0.80	0.98
45	-2.28	0.07	-0.80	0.98						
49	-2.28	0.07	-0.80	0.98						
52	-2.28	0.07	-0.80	0.98						
53	-2.28	0.07	-0.80	0.98						
54	-2.28	0.07	-0.80	0.98						

Hence, biology education students require a strong foundation in basic chemistry to solve problems related to biological phenomena effectively. However, 59 students have low problem-solving skills, while 29 students have very low problem-solving skills. Many biology education students struggle with basic chemistry concepts, detaining their ability to apply them to real-world biological problems (Wegner et al., 2019). Moreover, students with a better understanding of basic chemistry concepts and principles performed better in subsequent biology courses than those who lacked a solid understanding of basic chemistry (Mendez, et al., 2024). Integrating basic chemistry into biology education has a positive impact on students' learning outcomes, as it can enhance their understanding and retention of biological concepts (Shana & Abulibdeh, 2020). This is concerning, as problem-solving skills are essential for biology education students to develop and apply scientific knowledge to real-world issues in the classroom (Hsu et al., 2024). Teachers can employ strategies such as problem-based learning, case studies, and hands-on activities to promote problem-solving skills in basic chemistry. These strategies help students develop problem-solving skills and apply basic chemistry concepts to real-world biological problems.

Table 5. The response pattern of a misfit person

No	Student code	Response pattern from easy to difficult items					No	Student code	Response pattern from easy to difficult items				
		Item 5	Item 3	Item 2	Item 4	Item 1			Item 5	Item 3	Item 2	Item 4	Item 1
1	4	0	1	0	1	0	39	62	1	0	0	0	0
2	94	0	0	1	1	0	40	63	1	0	0	0	0
3	97	0	0	0	0	1	41	66	1	0	0	0	0
4	112	0	0	0	0	1	42	68	1	0	0	0	0
5	120	0	0	1	1	0	43	69	1	0	0	0	0
6	58	0	0	0	1	0	44	70	1	0	0	0	0
7	113	0	0	0	1	0	45	71	1	0	0	0	0
8	98	0	0	1	0	0	46	73	1	0	0	0	0
9	116	0	0	1	0	0	47	75	1	0	0	0	0
10	84	0	1	0	0	0	48	76	1	0	0	0	0
11	101	1	0	1	1	1	49	77	1	0	0	0	0
12	5	1	0	0	0	0	50	78	1	0	0	0	0
13	9	1	0	0	0	0	51	79	1	0	0	0	0
14	10	1	0	0	0	0	52	80	1	0	0	0	0
15	18	1	0	0	0	0	53	81	1	0	0	0	0
16	19	1	0	0	0	0	54	82	1	0	0	0	0
17	21	1	0	0	0	0	55	83	1	0	0	0	0
18	24	1	0	0	0	0	56	85	1	0	0	0	0
19	25	1	0	0	0	0	57	86	1	0	0	0	0
20	27	1	0	0	0	0	58	89	1	0	0	0	0
21	31	1	0	0	0	0	59	103	1	0	0	0	0
22	32	1	0	0	0	0	60	114	1	0	0	0	0
23	35	1	0	0	0	0	61	115	1	0	0	0	0
24	36	1	0	0	0	0	62	117	1	0	0	0	0
25	37	1	0	0	0	0	63	123	1	0	0	0	0
26	38	1	0	0	0	0							
27	40	1	0	0	0	0							
28	41	1	0	0	0	0							
29	42	1	0	0	0	0							
30	45	1	0	0	0	0							
31	49	1	0	0	0	0							
32	52	1	0	0	0	0							
33	53	1	0	0	0	0							
34	54	1	0	0	0	0							
35	55	1	0	0	0	0							
36	56	1	0	0	0	0							
37	60	1	0	0	0	0							
38	61	1	0	0	0	0							

*1: correct answer

0: incorrect answer

Using the Rasch model, we can analyze whether students provided responses that were appropriate to their level of knowledge or whether they cheated by copying answers from other students, based on the pattern of students' responses. The relationship between a student's ability and person misfit, where a student's response pattern differs from expected patterns based on their estimated ability level (misfitting response pattern), is significant for accurately interpreting test results. [Table 4](#) illustrates that 63 students are classified as misfit persons, with their abilities ranging from -2.29 logits to +2.24 logits. These students had varying abilities, ranging from very low to very high, and their response patterns could not be predicted by the Rasch Model ([Linacre, 2023](#); [Smith, 2001](#)). The misfitting response pattern did not significantly affect the item difficulty parameter, as deleting the misfitting responses resulted in only a negligible effect on the parameter's estimation ([Mousavi & Cui, 2020](#)). The misfitting response pattern is presented in [Table 5](#).

Person misfit often occurs when high-ability students answer more straightforward questions incorrectly, or when low-ability students answer difficult items correctly. It is caused by careless student responses, which may suggest random guessing, lack of engagement, or misunderstandings ([Patton et al., 2019](#); [Karabatsos, 2003](#)). Studies show that a person's misfit is often related to test-taking behavior, such as anxiety or fatigue,

which disrupts the alignment between ability and performance, leading to unreliable estimates of ability (Fox & Marianti, 2016). Therefore, a person's misfit can be caused by five factors: cheating, lucky guessing, careless responding, creative responding, and random responding (Goldammer et al., 2020; Karabatsos, 2003; Meijer, 1996). In Rasch models, person misfit can be detected through fit statistics, such as the infit and outfit indices, which measure unexpected response patterns in relation to item difficulty and student ability (Bond & Fox, 2015; Curtis, 2001).

For instance, if high-ability students consistently miss easier items, it could indicate poor item engagement or unclear item wording. In adaptive testing contexts, identifying misfits helps adjust future questions to improve accuracy and tailor assessments to individual student levels (Embretson & Reise, 2013). Addressing person misfits by identifying and mitigating factors that contribute to these irregularities, like unclear instructions or test anxiety, ultimately strengthens the validity of the assessment and the reliability of ability estimation. The instrument with an individual reliability of 0.07 and over 50% misfit is unsuitable for measuring individual abilities. Low Individual reliability indicates that the instrument fails to measure individual traits consistently; differences can make individual separation and reliability misleading indicators of measurement effectiveness (Linacre, 1995). Furthermore, the high misfit rates suggest aberrant response patterns, which compromise the validity of individual assessments (Sideridis & Jaffari, 2023). In this research, a person's misfit is caused by careless responding and creative responding factors. Students with high abilities can respond incorrectly to relatively easy questions because their way of thinking is unusual.

On the other hand, students with low skills can provide correct responses to complex questions because their way of thinking is more straightforward. Unpredictable response behavior can further complicate the interpretation of individual scores, reinforcing the notion that such instruments are more appropriate for exploratory diagnostics than definitive individual assessments (Turner & Engelhard, 2022). Thus, the combination of low reliability and significant discrepancies underscores the instrument's limitations in accurately measuring individual abilities (Braeken & van Laar, 2025). Therefore, the Rasch model can encounter problems due to the presence of both excessive uncertainty and excessive certainty in data analysis (Tesio et al., 2024). Thus, researchers must be careful when selecting samples so that the samples used provide representative results when tested with a research instrument.

The findings clearly demonstrate that all items of the problem-solving skills test in basic chemistry satisfy the requirements of the Rasch Model. The distribution of item difficulty indicates a well-calibrated range of cognitive demands, while the person-ability estimates reveal distinct and meaningful variation in students' problem-solving proficiency. However, this instrument demonstrates limited sensitivity in distinguishing between students' ability categories due to a low person-separation index. Therefore, these results suggest that the instruments are not designed for crucial or personalized decision-making, but instead for recognizing overall trends in students' problem-solving skills in fundamental chemistry. When analyzed at the group level, Rasch analysis helps uncover distributions of skill, item performance, and shared response patterns that can guide instructional reflection and enhance assessment improvements. Future studies should enhance person reliability by increasing the number of items, refining assessment criteria, and ensuring better alignment between item complexity and students' cognitive abilities. Moreover, integrating essay-type evaluations with supportive formats, such as structured responses or multi-tier items, could improve measurement accuracy while maintaining the evaluation of advanced cognitive skills. This research enhances the careful and transparent application of Rasch analysis to assess problem-solving abilities in the context of science education by recognizing these limitations and methodological factors.

CONCLUSION

Based on the result, the problem-based test is a valid and reliable tool for assessing students' problem-solving skills in basic chemistry. The findings indicate that most items demonstrate acceptable fit statistics, confirming that the instrument consistently measures students' problem-solving skills. The distribution of item difficulty reflects a balanced range of cognitive demands, while person-ability estimates reveal meaningful variation in students' problem-solving competence. Although the instrument performs adequately at the group level, the relatively low person separation index suggests limited sensitivity in distinguishing students into distinct high, medium, and low ability categories. Therefore, the instrument is more appropriate for identifying general trends in students' problem-solving skills than for high-stakes, individualized decisions. Further investigation is necessary to understand the characteristics of students who are misfitting the test, and to develop targeted interventions to support their learning. Teachers should consider using this test as a diagnostic tool to identify areas where students need additional support and scaffolding to improve their problem-solving skills in basic chemistry.

REFERENCES

- Abdellatif, H. (2023). Test Results With and Without Blue Printing: Psychometric Analysis Using the Rasch Model. *Educación Médica*, 24(3), 100802.
- Abo-Khalil, A. G. (2024). Integrating Sustainability into Higher Education Challenges and Opportunities for Universities Worldwide. *Heliyon*, 10(9), e29946.
- Adedoyin, O. O., & Mokobi, T. (2013). Using IRT Psychometric Analysis in Examining the Quality of Junior Certificate Mathematics Multiple Choice Examination Test Items. *International Journal of Asian Social Science*, 3(4), 992–1011.
- Alturki, U., & Aldraiweesh, A. (2023). The Factors Influencing 21st Century Skills and Problem-Solving Skills: The Acceptance of Blackboard as Sustainable Education. *Sustainability*, 15(17), 12845.
- Amanda, F. F., Sumitro, S. B., Lestari, S. R., & Ibrohim. (2024). Enhancing Critical Thinking and Problem Solving Skills by Complexity Science-Problem Based Learning Model. *Multidisciplinary Journal of Educational Research*, 14(1), 96–114.
- Artama, K. K., Budasi, I. G., & Ratminingsih, N. M. (2023). Promoting the 21st Century Skills Using Project-Based Learning. *Language Circle: Journal of Language and Literature*, 17(2), 325–332.
- Asyhari, A., & Sifa'i, M. (2021). Problem-Based Learning to Improve Problem Solving Skill: Is It Effective Enough? *Indonesian Journal of Science and Mathematics Education*, 4(1), 78–88.
- Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
- Balogun, A. G., Balogun, S. K., & Onyencho, C. V. (2017). Test Anxiety and Academic Performance Among Undergraduates: The Moderating Role of Achievement Motivation. *The Spanish Journal of Psychology*, 20(14), 1–8.
- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical Considerations for Using Exploratory Factor Analysis in Educational Research. *Practical Assessment, Research & Evaluation*, 18(6), 1–13.
- Biancardi, A., Colasante, A., D'Adamo, I., Daraio, C., Gastaldi, M., & Uricchio, A. F. (2023). Strategies for Developing Sustainable Communities in Higher Education Institutions. *Scientific Reports*, 13, 20596.
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research. *Frontiers in Public Health*, 6, 149.
- Bolbolian, M., Asgari, S., Sefidi, F., & Zadeh, A. S. (2021). The Relationship Between Test Anxiety and Academic Procrastination Among the Dental Students. *Journal of Education and Health Promotion*, 10, 1–6.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.
- Boran, M., & Karakus, F. (2022). The Mediator Role of Critical Thinking Disposition in the Relationship Between Perceived Problem-Solving Skills and Metacognitive Awareness of Gifted and Talent Students. *Participatory Educational Research*, 9(1), 61–72.
- Braeken, J., & van Laar, S. (2025). Reducing Calibration Bias for Person Fit Assessment by Mixture Model Expansion. *Educational and Psychological Measurement*, 85(1), 1–24.
- Bueno, P. M. (2014). Assessment of Achievement in Problem-Solving Skills in a General Chemistry. *Journal of Technology and Science Education*, 4(4), 260–269.
- Centoni, M., & Maruotti, A. (2021). Students' Evaluation of Academic Courses: An Exploratory Analysis to an Italian Case Study. *Studies in Educational Evaluation*, 70, 101054.
- Chi, L. T., & Ninh, T. T. (2018). Assessment of Problem Solving Ability and Creativity in Chemistry Teaching at Secondary Schools in Binh Dinh, Vietnam. *American Journal of Educational Research*, 6(6), 757–762.
- Contrino, M. F., Reyes-Millán, M., Vázquez-Villegas, P., & Membrillo-Hernández, J. (2024). Using an Adaptive Learning Tool to Improve Student Performance and Satisfaction in Online and Face-to-Face Education for a More Personalized Approach. *Smart Learning Environments*, 11, 6.

- Costales, J. A., Catulay, J. J., Bermudez, N. P., & Costales, J. A. (2022). Kaiser-Meyer-Olkin Factor Analysis: A Quantitative Approach on Mobile Gaming Addiction Using Random Forest Classifier. *Proceedings of the 6th International Conference on Information System and Data Mining* (pp. 18–24). Association for Computing Machinery.
- Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed.). Sage Publications.
- Curtis, D. D. (2001). Misfits: People and Their Problems. What Might It All Mean?. *International Education Journal*, 2(4), 91–99.
- Davis, D. R., & Boone, W. (2021). Using Rasch Analysis to Evaluate the Psychometric Functioning of the Other-Directed, Lighthearted, Intellectual, and Whimsical (OLIW) Adult Playfulness Scale. *International Journal of Educational Research Open*, 2, 100054.
- Demitra, & Sarjoko. (2018). Effects of Handep Cooperative Learning Based on Indigenous Knowledge on Mathematical Problem Solving Skill. *International Journal of Instruction*, 11(2), 103–114.
- Dikmen, M. (2023). Test Anxiety in Online Exams: Scale Development and Validity. *Current Psychology*, 42, 30210–30222.
- Doleck, T., Bazalais, P., Lemay, D. J., Saxena, A., & Basne, R. B. (2017). Algorithmic Thinking, Cooperativity, Creativity, Critical Thinking, and Problem Solving: Exploring the Relationship Between Computational Thinking Skills and Academic Performance. *Journal of Computers in Education*, 4(4), 355–369.
- Dünya, B. A. (2024). Adapting the Person Fit Analysis: Ideas on Detecting Person Misfit in Computerized Adaptive Testing. *Journal of Measurement and Evaluation in Education and Psychology*, 15(1), 1–4.
- Effatpanah, F., Baghaei, P., Ravand, H., & Kunina-Habenicht, O. (2024). Fitting the Mixed Rasch Model to the Listening Comprehension Section of the IELTS: Identifying Latent Class Differential Item Functioning. *International Journal of Testing*, 25(1), 50–89.
- Eidesen, P. B., Bjune, A. E., & Lang, S. I. (2023). "Show Me How to Use a Microscope" – The Development and Evaluation of Certification as Direct Assessment of Practical Lab Skills. *Ecology and Evolution*, 13(10), e10592.
- Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory for Psychologists*. Psychology Press.
- Fährmann, K., Köhler, C., Hartig, J., & Heine, J. H. (2022). Practical Significance of Item Misfit and Its Manifestations in Constructs Assessed in Large-Scale Studies. *Large-Scale Assessments in Education*, 10(7).
- Fatmawati, B., Marzuki, M., Roshayanti, F., & Suprpto, P. K. (2024). Fostering Students' Problem-Solving Skills Through Biology Learning Model Integrated With Kurikulum Merdeka. *Jurnal Pendidikan Biologi Indonesia*, 10(2), 392–403.
- Fauzie, M., Pada, A. U., & Supriatno. (2021). Analysis of the Difficulty Index of Item Bank According to Cognitive Aspects During the COVID-19 Pandemic. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(2), 162–169.
- Fischer, L., Rohm, T., Carstensen, C. H., & Gnamb, T. (2021). Linking of Rasch-Scaled Tests: Consequences of Limited Item Pools and Model Misfit. *Frontiers in Psychology*, 12, 633896.
- Fox, J. P., & Marianti, S. (2016). Joint Modeling of Ability and Differential Speed Using Responses and Response Times. *Multivariate Behavioral Research*, 51(4), 540–553.
- Fr'echette-Simard, C., Plante, I., Duchesne, S., & Chaffee, K. E. (2023). A Latent Growth Analysis of Individual Factors Predicting Test Anxiety During the Transition From Elementary to Secondary School. *Journal of Early Adolescence*, 43(3), 265–293.
- Gebremichael, M. W., Baraki, B., Mehari, A. B., & Assalfew, B. (2025). Item Analysis of Multiple Choice Questions From Assessment of Health Sciences Students, Tigray, Ethiopia. *BMC Medical Education*, 25.
- Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless Responding in Questionnaire Measures: Detection, Impact, and Remedies. *The Leadership Quarterly*, 31(4), 101384.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer Academic Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.

- Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, 9(2), 139–164.
- Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and How Do Students Learn?. *Educational Psychology Review*, 16, 235–266.
- Hrnjičić, A., & Alihodžić, A. (2024). Measuring Students' Conceptual Understanding of Real Functions: A Rasch Model Analysis. *International Electronic Journal of Mathematics Education*, 19(1), em0761.
- Hsu, J. L., Sung, R. J., Swarat, S. L., Gore, A. J., Kim, S., & Lo, S. M. (2024). Variations in Student Approaches to Problem Solving in Undergraduate Biology Education. *CBE—Life Sciences Education*, 23(2), 1–19.
- Jamshidi, S., Rezaei, S., Hassanzadeh, M., & Dehqan, M. (2019). Development and Validation of an Authorial Identity Model and Questionnaire: A Factor Analytic Approach. *Issues in Language Teaching*, 8(2), 243–273.
- Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education*, 16(4), 277–298.
- Karan, E., & Brown, L. (2022). Enhancing Student's Problem-Solving Skills Through Project-Based Learning. *Journal of Problem Based Learning in Higher Education*, 10(1), 74–87.
- Küçükaydın, M. A., & Ayaz, E. (2025). Validation of the Scientific Reasoning Competencies Instrument: Relationships With Epistemological Beliefs. *Instrument: Relationships with Epistemological Beliefs*, 23.
- Kusuma, I. Y., Triwibowo, D. N., Pratiwi, A. D., & Pitaloka, D. A. (2022). Rasch Modelling to Assess Psychometric Validation of the Knowledge About Tuberculosis Questionnaire (KATUB-Q) for the General Population in Indonesia. *International Journal of Environmental Research and Public Health*, 19(24), 16753.
- Lahza, H., Smith, T. G., & Khosravi, H. (2022). Beyond Item Analysis: Connecting Student Behaviour and Performance Using E-Assessment Logs. *British Journal of Educational Technology*, 54, 355–354.
- Lee, H. Y., Chen, Y. L., & Weng, L. J. (2023). Brief Research Report: Effects of Sampling Error and Categorization on Estimation of Measure of Sampling Adequacy. *The Journal of Experimental Education*, 92(2), 398–410.
- Lestari, S., & Yusuf, F. N. (2025). Aligning Assessment Practices With Learning Objectives: A Case of EFL Classes in Indonesia. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 10(1).
- Linacre, J. M. (1995). The Effect of Misfit on Measurement. *The Annual Meeting of the International Objective Measurement Workshop* (pp. 1–6).
- Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, 5(1).
- Linacre, J. M. (2022). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, 3(1).
- Linacre, J. M. (2023). *User Guide to Winsteps Ministep: Rasch-Model Computer Program*. Winsteps.
- Martin, T. J., Serrano-Estrada, L., & Esteve-Faubel, J. M. (2020). Critical Thinking: A Base for Urban Sustainable Development. *Higher Education Research & Development*, 40(2), 309–324.
- Meijer, R. R. (1996). The Influence of the Presence of Deviant Item Score Patterns on the Power of a Person-Fit Statistic. *Applied Psychological Measurement*, 20, 141–154.
- Mendez, L., Rivera, A. T., Vasquez, I., Aguilar, A. G., Owens, M. T., & Owens, C. L. (2024). How Students Taking Introductory Biology Experience the Chemistry Content. *Journal of Microbiology & Biology Education*, 25(3).
- Mousavi, A., & Cui, Y. (2020). The Effect of Person Misfit on Item Parameter Estimation and Classification Accuracy: A Simulation Study. *Education Sciences*, 10, 0324.
- Müller, M. (2020). Item Fit Statistics for Rasch Analysis: Can We Trust Them?. *Journal of Statistical Distribution and Applications*, 7(5).
- Nansen, C. (2024). Active Learning, Living Laboratories, Student Empowerment, and Urban Sustainability. *Sustainability*, 16(10), 3902.
- Nkansah, B. K. (2018). On the Kaiser-Meyer-Olkin's Measure of Sampling Adequacy. *Mathematical Theory and Modeling*, 8(7), 52–76.
- Nold, H. (2017). Using Critical Thinking Teaching Methods to Increase Student Success: An Action Research Project. *International Journal of Teaching and Learning in Higher Education*, 29(1), 17–32.

- Ntumi, S. (2025). Using Hierarchical Multiple Regression to Model the Impact of AI-Powered Adaptive Testing on Student Academic Fortunes and Test Anxiety in the Ghanaian Context. *Discover Education*, 141, 4.
- Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and Treatment of Careless Responses to Improve Item Parameter Estimation. *Journal of Educational and Behavioral Statistics*, 44(3), 309–341.
- Patyka, N., Gryschenko, O., Kucher, A., Hełdak, M., & Raszka, B. (2021). Assessment of the Degree of Factors Impact on Employment in Ukraine's Agriculture. *Sustainability*, 13(2), 564.
- Pitaloka, D. A., Kusuma, I. Y., Pratiwi, H., & Pradipta, I. S. (2023). Development and Validation of Assessment Instrument for the Perception and Attitude Toward Tuberculosis Among the General Population in Indonesia: A Rasch Analysis of Psychometric Properties. *Frontiers in Public Health*, 11, 1143120.
- 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, 020111.
- Polat, M. (2020). Analysis of Multiple-Choice Versus Open-Ended Questions in Language Tests According to Different Cognitive Domain Levels. *Novitas-ROYAL (Research on Youth and Language)*, 14(2), 76–96.
- Post, M. W., Fellinghuier, C. S., Charlifue, S., New, P. W., Forchheimer, M. B., & Tate, D. G. (2022). Rasch Analysis of the International Quality of Life Basic Data Set Version 2.0. *Archives of Physical Medicine and Rehabilitation*, 103, 2120–2130.
- Price, E., Nicholson, D. T., Dunk, R., Lawler, C., Carney, M., ... & Mottram, S. (2024). Enabling Change Agents for Sustainable Development: A Whole-Institution Approach to Embedding the UN Sustainable Development Goals in Higher Education. *International Journal of Sustainability in Higher Education*, 25(7), 1333–1350.
- Rad, H. S. (2025). Reinforcing L2 Reading Comprehension Through Artificial Intelligence Intervention: Refining Engagement to Foster Self-Regulated Learning. *Smart Learning Environments*, 12, 23.
- Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, 4(3), 207–230.
- Rezigalla, A. A., Eleragi, A. M., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., ... & Adam, M. I. (2024). Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items. *BMC Medical Education*, 24, 445.
- Riaz, S., Kousar, S., Saddique, R., & Rafiq, M. (2023). An Analysis of Laboratory Skills Among University Science Students. *iRASD Journal of Educational Research*, 4(1), 10–17.
- Riyadi, Syarifah, T. J., & Nikmaturrohman, P. (2021). Profile of Students' Problem-Solving Skills Viewed From Polya's Four Steps Approach and Elementary School Students. *European Journal of Educational Research*, 10(4), 1625–1638.
- Rodríguez-Zurita, D., Jaya-Montalvo, M., Moreira-Arboleda, J., Díez, E., & Carrión-Mero, P. (2025). Sustainable Development Through Service Learning and Community Engagement in Higher Education: A Systematic Literature Review. *International Journal of Sustainability in Higher Education*, 25(1), 158–201.
- Rosak-Szyrocka, J., & Knop, K. (2024). Sustainable Development Universities-Opportunities & Challenges on the Road to Society 5.0 From the Key Stakeholder Perspective. *PLoS One*, 19(8), e0308929.
- Sa-ngiemjit, M., Vázquez-Alonso, A., & Mas, M. A. (2024). Problem-Solving Skills of High School Students in Chemistry. *International Journal of Evaluation and Research in Education*, 13(3), 1825–1831.
- Shadle, S. E., Brown, E. C., Towns, M. H., & Warner, D. L. (2012). A Rubric for Assessing Students' Experimental Problem-Solving Ability. *Journal of Chemical Education*, 89(3), 319–325.
- Shana, Z., & Abulibdeh, E. S. (2020). Science Practical Work and Its Impact on Students' Science Achievement. *Journal of Technology and Science Education*, 10(2), 199–215.
- Sideridis, G., & Jaffari, F. (2023). Identifying Person Misfit Using the Person Backward Stepwise Reliability Curve (PBRC). *Frontiers in Psychology*, 14, 1273582.
- Sigudla, J., & Maritz, J. E. (2023). Exploratory Factor Analysis of Constructs Used for Investigating Research Uptake for Public Healthcare Practice and Policy in a Resource-Limited Setting, South Africa. *BMC Health Services Research*, 23, 1423.
- Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying Computerized Adaptive Testing to the CES-D Scale: A Simulation Study. *Psychiatry Research*, 188, 147–155.

- Soeharto, S., & Csapó, B. (2022). Exploring Indonesian Student Misconceptions in Science Concepts. *Heliyon*, 8, e10720.
- Souza, M. A., Coster, W. J., Mancini, M. C., Dutra, F. C., Kramer, J., & Sampaio, R. F. (2017). Rasch Analysis of the Participation Scale (P-Scale): Usefulness of the P-Scale to a Rehabilitation Services Network. *BMC Public Health*, 17, 1–9.
- Stehle, S. M., & Peters-Burton, E. E. (2019). Developing Student 21st Century Skills in Selected Exemplary Inclusive STEM High Schools. *International Journal of STEM Education*, 6(39).
- Sudjana, N. (2017). *Penilaian Hasil Proses Belajar Mengajar*. Remaja Rosdakarya.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata.
- Tan, J. P., Choo, S. S., Kang, T., & Liem, G. A. (2017). Educating for Twenty-First Century Competencies and Future-Ready Learners: Research Perspectives From Singapore. *Asia Pacific Journal of Education*, 37(4), 425–436.
- Tasgin, A., & Dilek, C. (2023). The Mediating Role of Critical Thinking Dispositions Between Secondary School Student's Self-Efficacy and Problem-Solving Skills. *Thinking Skills and Creativity*, 50, 101400.
- Tavakol, M., & Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, 2, 53–55.
- Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting Results From Rasch Analysis 2: Advanced Model Applications and the Data-Model Fit Assessment. *Disability and Rehabilitation*, 46(3), 04–617.
- Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J. M., Morisseau, T., ... & Lubart, T. (2023). Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *Journal of Intelligence*, 11(3), 54.
- Turner, K. T., & Engelhard, G. (2022). Using Functional Clustering to Diagnose Person Misfit. *The Journal of Experimental Education*, 92(2), 377–397.
- Underwood, S. M., Posey, L. A., Herrington, D. G., Carmel, J. H., & Cooper, M. M. (2018). Adapting Assessment Tasks to Support Three-Dimensional Learning. *Journal of Chemical Education*, 95, 207–217.
- Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, 44(3), 219–246.
- Wegner, C., Bruneman, R. L., & Steffen, A. (2019). The Thing About Chemistry: Investigating the Fundamental Chemistry Skills of Biology Education Students at Bielefeld University. *Science Education International*, 30(4), 334–341.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Mesa Press.
- Wu, Q., Zhang, Z., Song, Y., Zhang, Y., Zhang, F., & Li, R. (2013). The Development of Mathematical Test Based on Item Response Theory. *International Journal of Advancements in Computing Technology*, 5(10), 209–216.
- Yang, F. M., & Kao, S. T. (2014). Item Response Theory for Measurement Validity. *Shanghai Archives of Psychiatry*, 26(3), 171–177.
- Yasin, Y. M., Kehyayan, V., Khraim, F., & Al-Lenjawi, B. (2023). Psychometric Evaluation of the Acute Care Nurses' Job Satisfaction Scale-Revised. *Nursing Open*, 10, 488–497.
- Yerushalmi, E., & Eylon, B. (2015). Problem Solving in Science Learning. In R. Gunstone (Ed.), *Encyclopedia of Science Education* (pp. 786–790). Springer.
- Ying, Y., & Tiemann, R. (2024). Development of an Assessment Tool for Collaborative Problem-Solving Skills in Chemistry. *Disciplinary and Interdisciplinary Science Education Research*, 6, 25.
- Ziegler, M., & Hagemann, D. (2015). Testing the Unidimensionality of Items. *European Journal of Psychological Assessment*, 31(4), 231–237.
- Zubairi, A. M., & Kassim, N. L. (2006). Classical and Rasch Analyses of Dichotomously Scored Reading Comprehension Test Items. *Malaysian Journal of ELT Research*, 2, 1–20.