



## PSYCHOMETRIC VALIDATION OF AN ANALYTICAL SKILLS TEST IN PHYSICS USING THE RASCH MODEL

**Abstract.** *The accurate assessment of analytical thinking in physics, particularly in magnetism, poses substantial challenges due to the limitations of conventional tools in measuring higher-order cognitive skills.*

*This study aimed to validate an analytical skills test in physics, based on Bloom's Revised Taxonomy, with an emphasis on the dimensions of differentiating, organizing, and attributing within the subject of magnetism. The assessment instrument, comprising 15 multiple-choice items, was administered to a sample of 84 twelfth-grade students. A psychometric evaluation was conducted utilizing the Rasch Model, which included analyses of item fit, reliability, unidimensionality, and gender fairness through Differential Item Functioning (DIF). The results indicated strong internal consistency, with a Cronbach's alpha of .91, and high item reliability at .94. Furthermore, the item-person distribution showed a well-aligned outcome. Principal Component Analysis (PCA) confirmed the unidimensionality of the test, while the DIF analysis revealed several items exhibiting gender-related response patterns, thereby suggesting a need for refinement to achieve fairness. The findings endorse the application of this validated instrument for assessing higher-order analytical thinking in the field of physics and highlight the significance of equitable assessment practices within STEM education.*

**Keywords:** *analytical skills, Bloom's Taxonomy, physics assessment, psychometric validation, Rasch model*

**Nurussaniah Nurussaniah,  
Punaji Setyosari, Dedi Kuswandi,  
Saïda Ulfa**  
*State University of Malang, Indonesia*

**Nurussaniah Nurussaniah,  
Punaji Setyosari,  
Dedi Kuswandi,  
Saïda Ulfa**

### Introduction

Analytical skills are vital in physics education because they enable students to solve complex problems and think logically. These skills help break down intricate concepts into simpler parts, facilitating easier understanding and problem-solving. They allow students to break down problems, identify relationships between concepts, and develop solutions through logical reasoning (Can, 2020; Spencer et al., 2020; Susiloningsih et al., 2024). In magnetism, these skills are essential for understanding the interactions between electric currents and magnetic fields. Key topics include magnetic fields, magnetic forces on moving charges, forces on current-carrying wires, and electromagnetic induction (Giancoli, 2014; Halliday et al., 2013). These concepts are crucial for practical applications such as electric motors and generators. Strengthening analytical skills enhances the link between theory and real-world applications, promoting critical thinking and problem-solving in today's technology-driven world.

Even while analytical skills are essential to physics education, evaluating them, especially regarding magnetism, remains a difficult task. The deeper cognitive processes needed for higher-order thinking are not well captured by many of the evaluation instruments now in use, which disproportionately concentrate on rote memorization and basic understanding (Abdellatif & Abdel-Gawad, 2020; Sari et al., 2022). There is still a significant discrepancy between educational goals and the assessment tools used in practice (Kamid et al., 2022; Solé-Llussà et al., 2022), despite research and curricula highlighting the importance of Higher Order Thinking Skills (HOTS) (Kurniawan et al., 2022; Surahman et al., 2018). Attempts to develop advanced analytical skills are weakened by this mismatch, especially in magnetism, where students must combine theoretical concepts with real-world problem-solving techniques.

Studies repeatedly demonstrate that students with weak analytical abilities have difficulty handling difficult real-world situations as well as understanding complicated physics ideas (DeWitt et al., 2019; Ghanizadeh, 2017; Pullicino & Bonello, 2020; Tientongdee, 2018). The dearth of validated



evaluation tools created especially to gauge analytical abilities in magnetism exacerbates these challenges even more. Magnetism requires higher degrees of cognitive involvement than basic physics courses due to it incorporating abstract and complex ideas like electromagnetic induction and magnetic force interactions (Giancoli, 2014). These expectations are frequently not met by the assessment instruments that exist today, which leaves a gap in the ability to measure and promote higher-order thinking in this area.

Studies have also shown that gender disparities in physics performance may be attributable to variations in test format (Wheeler & Blanchard, 2019), content (Henderson et al., 2019), and cognitive demands (Kwarikunda et al., 2022). Consequently, it is imperative to ensure that assessment instruments are devoid of gender bias to mitigate the risk of misinterpreting students' competencies, which could impede educational equity. A valid instrument must accurately and equitably measure conceptual understanding (Kacprzyk et al., 2019). One established approach for detecting such bias is Differential Item Functioning (DIF) analysis, which assesses whether test items elicit differential responses from students of different genders despite equivalent proficiency levels (Boone et al., 2014). Therefore, scrutinizing gender-related bias is a fundamental step in the development of objective and equitable assessment instruments, particularly in physics education, where cognitive complexity varies across different conceptual domains.

The limited use of rigorous psychometric models, such as the Rasch Model, in physics examinations is another major obstacle. Rasch analysis has been frequently utilized to validate instruments in other fields (Tennant & Küçükdeveci, 2023), but it is still rarely used in physics education, especially when teaching subjects that call for higher-order analytical abilities. The majority of current research concentrates on broad cognitive skills or fails to evaluate certain Bloom's Taxonomy dimensions, such as differentiating, organizing, and attributing, which are critical for assessing physics analytical thinking (Damayanti et al., 2020; Ridhoni et al., 2023). In order to close this gap, assessment tools must be aligned with advanced learning objectives and validated using strong statistical techniques to guarantee accuracy, equity, and applicability for a range of student demographics.

Creating effective and trustworthy assessment instruments is crucial in order to precisely measure physics students' analytical abilities. While reliability ensures consistent outcomes across many situations, validity ensures that an instrument assesses the intended cognitive processes. Item validity, difficulty levels, and student ability distributions can all be effectively analyzed using the Rasch Model, a quantitative method within Item Response Theory (IRT) (Boone et al., 2014). Visual aids such as Person-Item Maps are also useful for determining whether an item's complexity and student proficiency match, as well as for pointing out items that need to be revised.

The Rasch Model is well-known and widely used in other fields, although it is still rarely used to assess analytical abilities in magnetism. There is a significant lack of validated instruments created especially for secondary-level magnetism teaching since the majority of research concentrates on other physics subjects or stresses more general cognitive abilities. A thorough examination of analytical abilities in magnetism requires the inclusion of Bloom's Taxonomy dimensions (Anderson & Krathwohl, 2001), which include differentiating, organizing, and attributing, in assessment instruments.

### *Analytical Skill in Bloom's Taxonomy*

Cognitive processes in Bloom's Revised Taxonomy are categorized into six hierarchical levels: remembering (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6) (Anderson & Krathwohl, 2001). These levels are grouped into Lower-Order Thinking Skills (LOTS) for C1–C3 and Higher-Order Thinking Skills (HOTS) for C4–C6. In physics education, particularly in complex areas like magnetism, HOTS are crucial for grasping intricate concepts (Alanazi et al., 2024; Dessie et al., 2023; Maulana et al., 2025; Opfermann et al., 2017). Mastering topics such as magnetic field interactions and electromagnetic induction requires cognitive engagement that extends beyond simple memorization (Giancoli, 2014; Halliday et al., 2013).

This study specifically examines the analyzing domain (C4), which plays a vital role in enhancing students' problem-solving and scientific reasoning abilities. The analyzing domain comprises three sub-dimensions: differentiating, organizing, and attributing (Anderson & Krathwohl, 2001; Krathwohl, 2002). Differentiating involves distinguishing relevant information from irrelevant details, organizing focuses on structuring ideas and recognizing relationships, and attributing identifies underlying causes or principles based on evidence. These sub-dimensions are essential as they mirror the analytical thinking required in secondary physics instruction. For instance, students analyzing the motion of charged particles or explaining induced electromotive force must employ differentiating, organizing, and attributing skills.



Although several action verbs such as compare, contrast, infer, outline, and diagram are frequently associated with the analyzing process in instructional practice, it is important to note that these verbs are more appropriately understood as representations of observable learning behaviors within educational contexts, rather than as formal subcategories explicitly defined in the revised taxonomy framework. These learning outcome verbs serve as common pedagogical tools used to articulate instructional objectives and design assessment instruments; however, they do not directly reflect the theoretical constructs embedded in the official structure of Bloom's Revised Taxonomy (Churches, 2008; Forehand, 2010). This study adheres to the formal structure of the analyzing domain to maintain construct validity in assessment. Therefore, although many instructional verbs are associated with analysis in practice, this study adopts only the three sub-dimensions officially defined in Anderson and Krathwohl's revised taxonomy, differentiating, organizing, and attributing, to ensure conceptual consistency and construct validity in assessing analytical thinking.

### *Research Problem*

Despite the pivotal role of analytical competencies in physics education, particularly within the domain of magnetism, extant assessment instruments frequently fail to encapsulate the sophisticated cognitive processes requisite for higher-order thinking. Predominantly, contemporary evaluative tools emphasize rote recall rather than fostering a deep conceptual grasp, thereby constraining their efficacy in appraising students' problem-solving proficiencies. Furthermore, the potential presence of gender bias within these assessment frameworks may further obfuscate an accurate measurement of student competencies, engendering disparities in educational outcomes. Additionally, the limited integration of robust psychometric models, such as the Rasch Model, in physics assessments exacerbates these challenges, underscoring the need for the development of valid, reliable, and equitable evaluative instruments.

### *Research Focus*

This study thoroughly validates an assessment tool for analytical skills in the area of magnetism, utilizing Bloom's Taxonomy framework with a focus on the dimensions of differentiating, organizing, and attributing. The objective was to provide a comprehensive evaluation of students' cognitive skills while also exploring possible gender biases to ensure fairness and equity in the assessment process. The Rasch Model is employed for validating items, analyzing difficulty levels, and assessing the distribution of student abilities, thus guaranteeing that the instrument maintains psychometric soundness and delivers accurate, unbiased measurements.

### *Research Aim and Research Questions*

The research aimed to validate an analytical skills assessment instrument based on the Rasch model, specifically designed to align with the cognitive complexity of concepts related to magnetism. By employing psychometric evaluation and examining potential gender-related differences in item performance, this study sought to create a robust and equitable tool for measuring students' higher-order thinking skills in physics education. To achieve this aim, the study addressed four key research questions.

1. To what extent is the Rasch-based analytical skills test valid and reliable in the context of physics education?
2. How do inter-item relationships within Bloom's Taxonomy dimensions contribute to the construct validity of the instrument?
3. Does the distribution of item difficulty effectively reflect the full range of student abilities?
4. Is there evidence of gender-related bias in the instrument as identified through the DIF analysis?

## **Research Methodology**

### *General Background*

This research assessed the validity and reliability of tests designed to measure students' analytical skills. The instrument was constructed around the analytical dimensions of Bloom's Taxonomy, specifically differentiating, organizing, and attributing, to evaluate higher-order thinking skills (HOTS) in physics education. Conducted from



September to November 2024, the research followed a structured process involving test development, empirical data collection, and validation through Rasch modeling to ensure robust measurement integrity (Boone et al., 2014; Oreondo & Antonio, 1984).

### *Participants*

This study employed a purposive sample of 84 Grade 12 science-track students from public upper-secondary schools in West Kalimantan, Indonesia. Participants were selected from a defined population of students who had chosen physics as a specialization subject, thereby ensuring sufficient prior exposure to magnetism-related concepts. At the time of data collection, the implementation of specialization-based curricula varied across schools, with not all institutions offering physics as a dedicated track. This curricular inconsistency significantly limited the pool of eligible participants, reinforcing the rationale for focusing on a well-defined and pedagogically appropriate subgroup. Although the sample size may be considered moderate, it aligns with established guidelines recommending 50 to 100 participants for stable item calibration in Rasch analysis when applied to instruments with fewer than 20 items (Linacre, 1994). Moreover, as the primary objective of the study was to validate the psychometric properties of the instrument rather than to generalize findings, the use of a targeted, homogeneous sample was methodologically appropriate. Rasch modeling prioritizes the quality of item–person interactions over sheer sample size, and prior research supports the reliability of parameter estimates derived from academically uniform groups (Bond & Fox, 2015; Boone et al., 2014). Ethical clearance was obtained from the Research Ethics Committee of West Kalimantan, and informed consent was secured from all participants and, where applicable, their legal guardians.

### *Instrument and Procedures*

Prior to attempting the test, students are expected to understand key concepts related to magnetism, including the direction and structure of magnetic fields, magnetic forces on moving charges and current-carrying wires, solenoids, and electromagnetic induction. These concepts form the foundational knowledge needed to engage with the test items and were explicitly used to construct each question.

The analytical skills test comprised 15 items specifically designed to align with the three dimensions of Bloom's Taxonomy, as outlined in Table 1. Each item was developed based on those key physics concepts and mapped to one of the analyzing sub-dimensions: differentiating, organizing, or attributing. The corresponding concept targeted by each item is listed in the "Physics Concept" column to support content validity. The complete set of items is provided in the Appendix. Of the 15 items, 5 assess differentiating, 8 assess organizing, and 2 assess attributing. This distribution was informed by a cognitive analysis of secondary-level magnetism content, which indicates that instructional tasks more frequently engage students in distinguishing relevant information and organizing conceptual structures. In contrast, attributes focused on identifying causal relationships tend to appear less frequently and in more advanced reasoning contexts. Therefore, the allocation of items reflects the cognitive profile of the subject matter, ensuring proportional and meaningful representation of each sub-dimension.

**Table 1**  
*Alignment of Analytical Skills Test Items with Bloom's Taxonomy Dimensions*

| Item No. | Question focus                                                 | Cognitive process                                   | Analytical dimensions | Physics concept                               |
|----------|----------------------------------------------------------------|-----------------------------------------------------|-----------------------|-----------------------------------------------|
| S1       | Identifying the strongest magnetic force location              | Identifying relevant data from magnetic field lines | Differentiating       | Magnetic field strength and lines             |
| S2       | Determining electron trajectory in a uniform magnetic field    | Recognizing differences in motion paths             | Differentiating       | Lorentz force                                 |
| S3       | Identifying correct right-hand rule application                | Choosing the most appropriate force direction       | Differentiating       | Right-hand rule                               |
| S4       | Identifying charge types based on particle motion in a field   | Differentiating charge effects on trajectory        | Differentiating       | Motion of charged particles in magnetic field |
| S5       | Determining current direction from surrounding magnetic fields | Relating two variables to derive a conclusion       | Organizing            | Magnetic field around a straight wire         |

| Item No. | Question focus                                                             | Cognitive process                                     | Analytical dimensions | Physics concept                                |
|----------|----------------------------------------------------------------------------|-------------------------------------------------------|-----------------------|------------------------------------------------|
| S6       | Analyzing distance effect on force between parallel wires                  | Relating two variables to derive a conclusion         | Organizing            | Magnetic force between parallel conductors     |
| S7       | Interpreting tabulated data on magnetic force relationships                | Structuring knowledge in hierarchical relationships   | Organizing            | Relationship between current, field, and force |
| S8       | Predicting the effect of current increase on motor speed                   | Breaking down relationships between variables         | Organizing            | The motor effect                               |
| S9       | Determining the direction of induced current in a moving conductor         | Applying principles systematically                    | Organizing            | Lenz' law                                      |
| S10      | Evaluating the effect of coil windings and current on field strength       | Identifying factors affecting field intensity         | Attributing           | Magnetic field in solenoid                     |
| S11      | Differentiating the effects of current and coil windings on field strength | Distinguishing between two independent influences     | Differentiating       | Factor affecting solenoid field                |
| S12      | Ordering magnetic field strengths for different solenoids                  | Arranging relationships in logical order              | Organizing            | Solenoid design and magnetic field             |
| S13      | Calculating induced EMF based on coil windings and magnet motion           | Connecting motion parameters to voltage induction     | Organizing            | Faraday law                                    |
| S14      | Identifying the primary cause of induced EMF in a coil                     | Determining the key variable influencing a phenomenon | Attributing           | Cause of induced EMF                           |
| S15      | Estimating average induced EMF from a graphical dataset                    | Extracting information from visual representations    | Organizing            | Induced EMF and magnetic flux change           |

The test was administered in a controlled classroom environment to ensure consistency in data collection. Students were given 90 minutes to complete the test, which consisted of multiple-choice questions accompanied by contextual diagrams. The teacher provided instructions and monitored the session to ensure that all participants completed the test independently. Student responses were recorded and scored using a dichotomous format, where correct answers were assigned a score of 1 and incorrect answers a score of 0.

### *Data Analysis*

The collected data were analyzed using the Rasch Model to evaluate the psychometric properties of the instrument. Validity and reliability were assessed through multiple statistical measures, including Infit and Outfit Mean Square (MNSQ) statistics, with acceptable values ranging between 0.5 and 1.5, and person and item reliability indices to ensure consistency. A unidimensionality test using Principal Component Analysis (PCA) on Rasch residuals confirmed that the instrument measured a single construct. To examine internal consistency, item correlations within Bloom's Taxonomy dimensions were analyzed, and the distribution of items was reviewed for proportional representation and content coverage. The item difficulty distribution was evaluated through a Wright Map to determine how well test items aligned with student abilities, identifying any gaps where specific ability groups might be underrepresented. Additionally, Differential Item Functioning (DIF) analysis was conducted to detect potential gender-related bias, ensuring that no items were systematically easier or more difficult for specific gender groups. Items showing significant DIF were further examined and revised to improve measurement fairness and instrument accuracy.

## **Research Results**

### *Evaluation of Validity and Reliability in Analytical Skills Test*

To ensure the validity and reliability of the test, various psychometric properties were analyzed. Table 2 presents a comprehensive summary of the key psychometric properties of the test instrument.

**Table 2***Summary of Psychometric Properties*

| Category                      | Value                                   |
|-------------------------------|-----------------------------------------|
| Number of Students            | 84 (75 non-extreme)                     |
| Number of Items               | 15 (all non-extreme)                    |
| Student Ability Range (Logit) | -2.51 to +4.77                          |
| Item Difficulty Range (Logit) | -2.11 to +2.94                          |
| Mean Ability (Logit)          | 1.01 (non-extreme), 1.41 (all students) |
| Person Reliability            | .83 (non-extreme), .80 (all students)   |
| Item Reliability              | .94                                     |
| Cronbach's Alpha (KR-20)      | .91                                     |
| Separation Index              | 2.21 (students), 3.95 (items)           |
| Infit MNSQ (Mean)             | 0.95 (students), 0.97 (items)           |
| Outfit MNSQ (Mean)            | 1.09 (students & items)                 |
| Problematic Items             | 1 item with Outfit MNSQ = 4.56          |
| Extreme Students              | 9 students (10.7%)                      |

The dataset contains 84 student responses, where 75 non-extreme students receive distribution through the logit scale. The examination consists of 15 test items which fall under the non-extreme category. Student scores span between Logit -2.51 and +4.77 at the same time, item difficulties exist between Logit -2.11 and +2.94, which provides a balanced assessment. The measurement instrument shows robust reliability through a Person Reliability value of .83 for non-extreme students and a Cronbach's Alpha (KR-20) value of .91 for excellent internal consistency. The student ability levels display effective differentiation through the Separation Index of 2.21, and item calibration effectiveness is verified through an Item Reliability score of .94 and Separation Index at 3.95. A single item displays an Outfit MNSQ value of 4.56, signalling potential response issues possibly needing additional assessment. Further assessment items should be considered because 10.7% of students fall into the extreme range. The implemented measurement method shows effectiveness for determining middle-range student competency levels alongside strong measurement properties.

Principal Component Analysis (PCA) of Rasch residuals was performed in order to establish the unidimensionality of the instrument. Table 3 presents the results, indicating that the primary dimension accounts for 41.2% of the total variance, while the first and second contrasts contribute only 8.3% and 6.1%, respectively. These findings provide empirical support for the instrument's ability to measure a singular, coherent construct, with minimal evidence of multidimensionality.

**Table 3***PCA of Rasch Residuals*

| Component         | Explained Variance (%) |
|-------------------|------------------------|
| Primary Dimension | 41.2                   |
| First Contrast    | 8.3                    |
| Second Contrast   | 6.1                    |

### *Correlation Analysis of Subdimensions*

To further examine the internal structure of the instrument, a correlation analysis was conducted among items categorized within the three subdimensions of Bloom's Taxonomy: Differentiating, Organizing, and Attributing. The results, as presented in Table 4, indicate moderate and statistically significant correlations ( $r = .42-.60$ ,  $p < .001$ ) across item pairs within each subdimension.

**Table 4***Correlation Analysis of Subdimensions*

| Item Pair | Correlation ( <i>r</i> ) | <i>p</i> (2-tailed) | Dimension       |
|-----------|--------------------------|---------------------|-----------------|
| S1 - S3   | .50                      | < .001              | Differentiating |
| S2 - S11  | .46                      | < .001              | Differentiating |
| S5 - S7   | .48                      | < .001              | Organizing      |
| S15 - S9  | .60                      | < .001              | Organizing      |
| S10 - S14 | .42                      | < .001              | Attributing     |

The pattern of correlations observed across item pairs within each subdimension suggests a coherent internal structure, with items consistently measuring the targeted cognitive constructs. For example, items S1 and S3 ( $r = .50$ ) within the Differentiating subdimension assess students' ability to isolate relevant information, while S9 and S15 ( $r = .60$ ) in the Organizing subdimension reflect their capacity to synthesize and relate conceptual elements. Likewise, S10 and S14 ( $r = .42$ ) evaluate causal reasoning in Attributing. The moderate correlation coefficients underscore two key aspects of validity, demonstrating internal consistency within each subdimension and revealing conceptual interconnections that characterize higher-order analytical reasoning. Importantly, these associations are not excessively high, which supports the discriminant validity of the constructs and suggests that each subdimension captures a distinct analytical process. The combined findings from PCA and correlation analysis substantiate the instrument's validity in assessing analytical ability. While PCA confirms unidimensionality at the overall construct level, the correlation analysis highlights the cohesion among items within subdimensions, demonstrating the instrument's ability to capture distinct yet interconnected facets of higher-order analytical thinking.

*Item Difficulty Distribution and Wright Map Analysis*

Table 5 summarizes the ranking of test items based on difficulty levels, measured in logits. The easiest item (S8, -2.11 logits) and the most difficult item (S3, +2.94 logits) indicate a wide range of difficulty levels, effectively covering mid-range student abilities.

**Table 5***Item Fit Order*

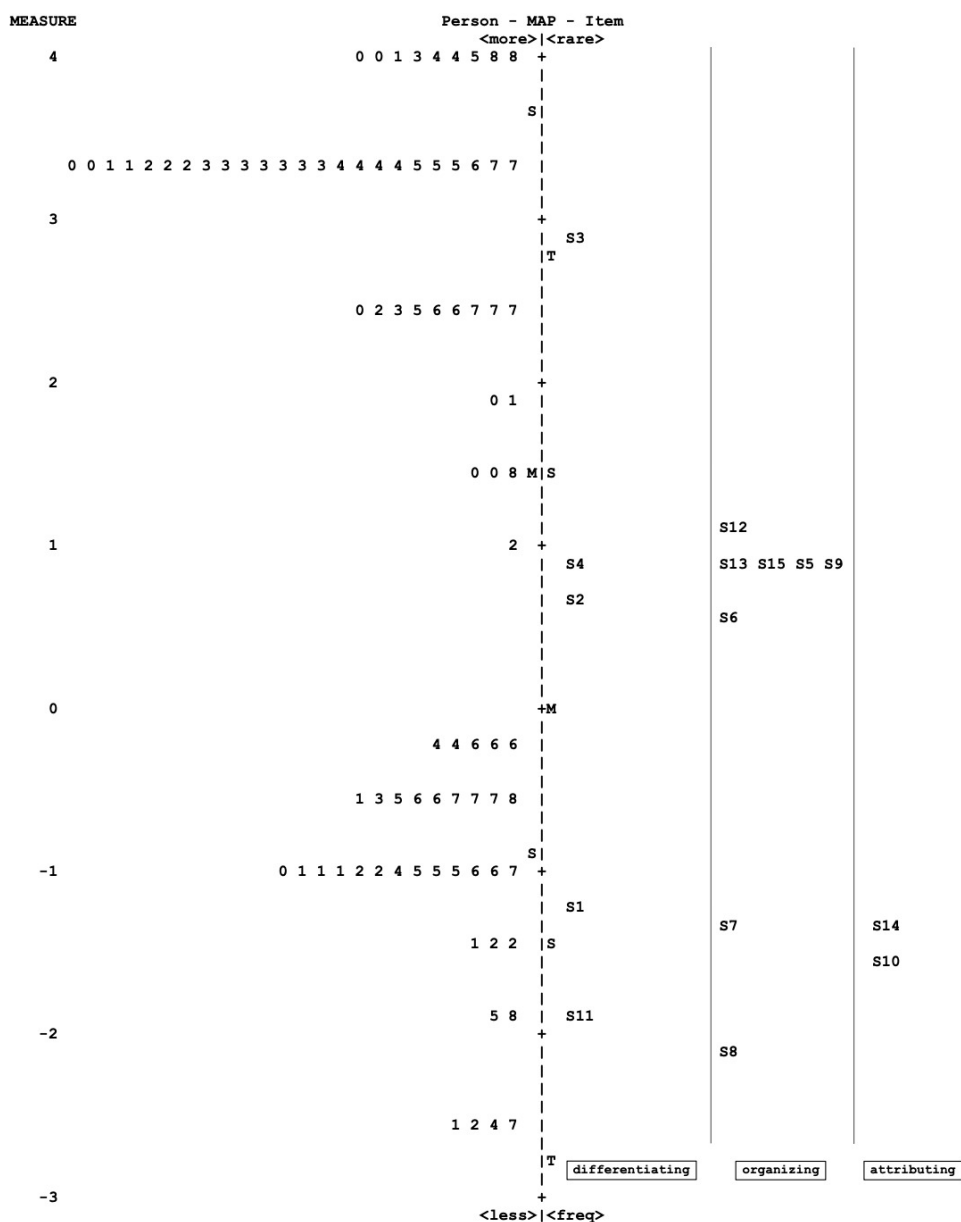
| Item | Difficulty (Logit) | Infit MNSQ | Outfit MNSQ | Remarks                                  |
|------|--------------------|------------|-------------|------------------------------------------|
| S8   | -2.11              | 0.75       | 0.68        | Easiest item, good fit                   |
| S1   | -1.23              | 1.29       | 1.19        | Borderline acceptable                    |
| S7   | -0.92              | 0.85       | 0.78        | Good fit                                 |
| S4   | -0.85              | 0.92       | 0.89        | Good fit                                 |
| S14  | -0.38              | 0.95       | 0.93        | Good fit                                 |
| S5   | -0.23              | 1.02       | 1.05        | Good fit                                 |
| S9   | -0.12              | 1.10       | 1.15        | Good fit                                 |
| S15  | +0.12              | 0.88       | 0.85        | Good fit                                 |
| S11  | +0.22              | 0.91       | 0.88        | Good fit                                 |
| S12  | +0.45              | 1.15       | 1.11        | Good fit                                 |
| S6   | +0.51              | 1.38       | 1.31        | Acceptable but high variation            |
| S2   | +0.62              | 2.17       | 2.01        | Misfit (requires revision)               |
| S10  | +1.35              | 1.24       | 1.21        | Acceptable                               |
| S13  | +1.90              | 0.98       | 1.02        | Good fit                                 |
| S3   | +2.94              | 1.61       | 4.56        | Hardest item, misfit (requires revision) |



Most items exhibit a good fit to the Rasch Model, with Infit MNSQ and Outfit MNSQ values falling within the acceptable range of 0.5 to 1.5, indicating consistent response patterns. Notably, S7, S4, S14, S5, and S9 demonstrate strong psychometric properties and are well-calibrated for assessing mid-range abilities. However, Item S2 (Outfit MNSQ = 2.01) and Item S3 (Outfit MNSQ = 4.56) display problematic fit statistics, suggesting irregular response patterns that may arise from ambiguous wording or an inappropriate difficulty level. While the overall item distribution aligns well with student abilities, ensuring comprehensive coverage, the inclusion of additional items targeting extreme ability groups could further enhance measurement precision. Refining misfitting items such as S2 and S3 is essential to improve model adherence and maintain the instrument's psychometric robustness.

The Wright Map (see Figure 1) illustrates the alignment between student abilities and item difficulties, confirming that the test effectively assesses mid-range abilities, though adjustments are needed to improve coverage of extreme student groups.

**Figure 1**  
*Wright Map Analysis*



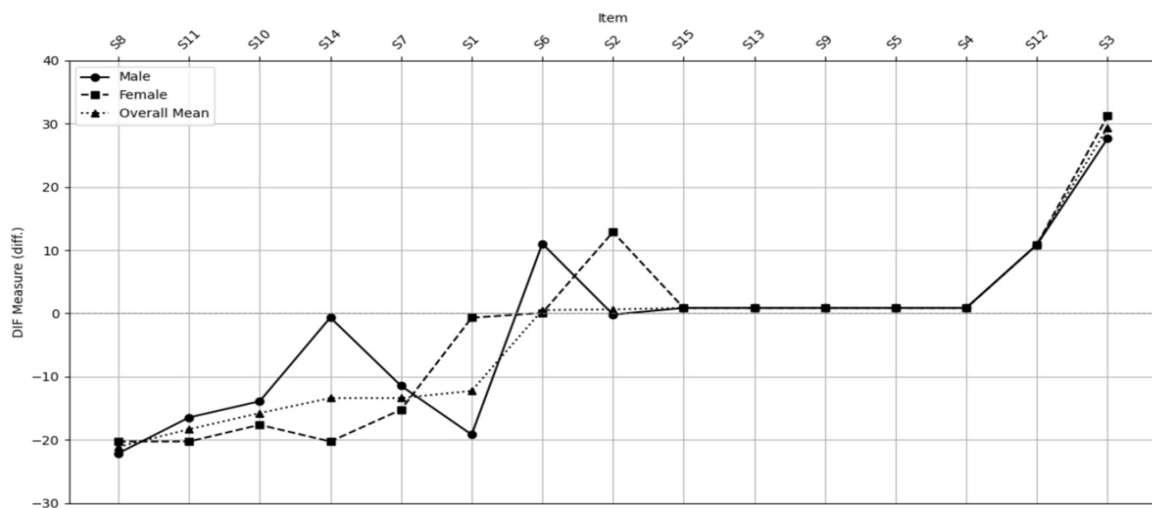
Three Bloom's Taxonomy analytical dimensions, named Differentiating, Organizing, and Attributing, categorize test items to assess different cognitive functions, including difference recognition, information organization, and causal relationship interpretation. The Differentiating assessment components (S1, S2, S3, S4, S11) measure students' ability to discern important differences and select essential information while spanning a difficulty scale from -2 to +3 logits. The questions S1 and S11 (-2 logits) support lower-ability students, but S2 and S4 (0 to +1 logits) target average-level learners, and S3 (+3 logits) present significant difficulty requiring revision to become accessible. The Organizing section (S5, S6, S7, S8, S9, S12, S13, S15) evaluates student capabilities to structure and integrate information through items scoring between 0 to +2 logits, which effectively addresses mid-level cognitive processes. The basic structuring skills are evaluated through S5 and S6, which are both at 0 logits, yet S12 and S15 test advanced synthesis abilities with +2 logits. The lack of simpler difficulty-level items reveals an important need to develop additional straightforward items that will improve accessibility for all students. The Attribute subdomain (S10, S14) explores causal relationships through items rated from -1 logit for basic causal connections to +2 logits for analyzing multiple interacting variables. This dimension shows effective ability coverage but fails to include challenging items for complex causal reasoning at the advanced level. The Organizing dimension controls the assessment by measuring middle-level abilities, yet Differentiating and Attributing dimensions expand the scope throughout basic and advanced cognitive processes. Most items in the test range between 0 to +2 logits in difficulty level, which matches the ability range of students. Additional modifications to S3, together with new low-difficulty items, should be implemented to improve both the accuracy and inclusiveness of this instrument.

#### *Differential Item Functioning (DIF) Analysis*

To further investigate potential gender-related biases in item difficulty, a Differential Item Functioning (DIF) analysis was conducted. The results are illustrated in Figure 2, which presents a Person DIF Plot comparing DIF measuring across respondent groups (males, females, and overall averages). This plot allows for an evaluation of whether certain items systematically favor one group over another.

**Figure 2**

*Person DIF Plot Comparing DIF Measures for Each Test Item Across Gender Groups*



The horizontal X-axis shows item codes like S8, S11, and S10, while the vertical Y-axis displays DIF Measure values representing item difficulty inconsistencies between groups. A positive DIF Measure shows an item appears easier for one particular group, yet a negative DIF Measure demonstrates the item becomes more difficult for that group. Male responses are represented by solid lines with black circles, female responses by dashed lines with black squares, and the overall mean by dotted lines with black triangles.



The data analysis identified S14, S1, S6, and S2 as items demonstrating substantial differences in performance levels between male and female participants. Item S14 finds females face elevated challenges compared to males, yet items S1 and S2 show males face more obstacles than females. Item S6 reveals a significant discrepancy as female survey participants achieved lower scores than their male counterparts. The items S11, S10, S7, and S3 present moderate variations among respondents, yet maintain acceptable performance ranges. The differential item functioning (DIF) values for S12, S15, S13, S9, S5, and S4 approach zero indicate that male and female survey participants demonstrate comparable performance when answering these specific items.

Variations in item difficulty indicate that specific group-related factors influence measurement complexity. These differences may arise from distinct cognitive styles, varying learning experiences, or linguistic factors that advantage certain groups. Therefore, it is essential to conduct a more comprehensive investigation of these items to identify potential gender-related biases, whether in question formulation or evaluation processes. A more comprehensive evaluation of this ability test will yield more precise and objective assessments of student performance by minimizing unintended biases.

## Discussion

### *Validity and Reliability of the Rasch-Based Analytical Skills Test in Physics Education*

This study evaluated the validity and reliability of a Rasch-based test instrument designed to assess students' analytical skills in the context of magnetism. The findings confirm the instrument's robust psychometric properties, while also identifying areas for improvement to enhance inclusivity and fairness. Bloom's Revised Taxonomy designates analyzing (C4) as the pivotal transition to higher-order cognition, comprising the sub-processes of differentiating, organizing, and attributing (Anderson & Krathwohl, 2001). The present instrument systematically aligns each item with these three sub-processes, thereby ensuring that the observed indices of validity and reliability are not incidental artifacts but empirical evidence that Bloom's theoretical constructs can be rigorously quantified through Rasch modeling. Consequently, the findings substantiate the theoretical proposition that the C4 domain is both measurable and divisible into discrete cognitive components.

The instrument demonstrated strong reliability, with a Cronbach's alpha of .91 and item reliability of .94, indicating consistent measurement of analytical skills. These values align with Rasch model standards, where reliability indices above .80 are considered strong for both person and item reliability (Bond & Fox, 2015; Boone et al., 2014).

The separation values further support the instrument's discriminative capability, with 4.30 for items and 2.21 for persons, ensuring that the test can effectively classify students into distinct ability levels. A separation value above 2.0 is widely recognized as a threshold for good discrimination (Yuhanna et al., 2021), while values above 3.0 for items and 2.0 for persons indicate optimal measurement precision (Jamar & Siew, 2020). Additionally, a person's separation value above 2.5 confirms a good model fit, reinforcing the test's sensitivity in distinguishing varied ability levels (Alhadabi & Aldhafri, 2021; Chen et al., 2021).

The PCA of Rasch residuals validates the unidimensionality of the instrument, with the primary dimension accounting for 41.2% of the variance, while the first contrast explains only 8.3%. A variance explanation above 40% confirms an instrument's coherence in measuring a single construct, as emphasized in prior research (Denovan et al., 2022, 2024). Additionally, the low residual variance in the first contrast suggests minimal secondary dimensions, further strengthening the instrument's validity in assessing analytical skills.

### *Alignment of Items and Student Abilities*

The Wright Map Analysis demonstrates a strong alignment between item difficulties and student abilities, particularly within the 0 to +2 logit range, indicating that the instrument effectively assesses average and above-average students. This finding aligns with previous research, which emphasizes that matching assessment items to the ability spectrum is essential for optimizing measurement accuracy and ensuring that tests remain appropriately challenging and fair across diverse student backgrounds (Krell et al., 2021; Mancuso et al., 2019; Tavares-Júnior et al., 2019).

However, gaps at the extremes of the ability spectrum highlight the need for a more comprehensive assessment that addresses students with low abilities (< -2 logits) and high abilities (> +3 logits). Assessments often

lack adequate representation for extreme ability levels, calling for more inclusive designs (Lestari & Samsudin, 2020). Similarly, aligning test items with student competencies and learning outcomes is essential to ensure that all students, including those at the highest and lowest ends of the spectrum, are appropriately challenged and supported in reaching their full potential (Chan et al., 2023).

The misalignment is particularly evident in S8 (-2.11 logits) and S3 (+2.94 logits). While S8 is accessible to low-ability students, S3 presents a significant challenge, with substantial response misfit (Outfit MNSQ = 4.56). These findings suggest that while the instrument is effective for mid-range abilities, additional items targeting extreme ability levels are necessary to enhance assessment inclusivity and precision.

#### *Construct Validity and Inter-Item Relationships in Bloom's Taxonomy*

The correlation analysis within Bloom's Taxonomy dimensions reveals moderate to strong relationships and statistically significant correlations ( $r = .42-.60$ ,  $p < .001$ ), supporting the alignment of items with their intended cognitive processes. This finding is consistent with previous research, which reported an intraclass correlation coefficient (ICC) of .67, indicating moderate correlation among items and reinforcing acceptable test-retest reproducibility (Diniz et al., 2019). These results demonstrate that moderate correlations not only validate item alignment with cognitive processes but also contribute to instrument reliability, strengthening its validity for consistent use in assessments.

Differentiating items, spanning a difficulty range from -1 to +3 logits, provide a comprehensive assessment from foundational to advanced analytical skills. Distinguishing between relevant and irrelevant information is fundamental to analytical thinking, as emphasized in previous studies (Darmaji et al., 2019; Kamid et al., 2022).

Organizing items, concentrated around the mean difficulty (0 to +2 logits), are well-calibrated for assessing average student abilities. Tasks with moderate difficulty effectively evaluate students' cognitive abilities by encouraging systematic information integration (Schwippert et al., 2024). Well-structured assessments at average difficulty levels foster meaningful engagement and ensure valid measurement of analytical thinking without introducing unnecessary complexity (Moussa, 2021). Additionally, systematically organizing assessments enhances the accurate reflection of students' mid-level cognitive skills and promotes effective learning outcomes (Kliziene et al., 2022).

Attributing items, which focus on deconstructing causal relationships, exhibit a balanced difficulty range (-1 to +2 logits), effectively assessing higher-order thinking skills. Such calibration enhances students' abilities to analyze causal relationships in physics, fostering deeper conceptual engagement (Ramadhan et al., 2019). Balancing difficulty levels is essential for accurately evaluating students' analytical and synthesis skills (Agustihana & Suparno, 2019), while well-structured items within this range facilitate a comprehensive assessment of cognitive abilities (Erfianti et al., 2019).

#### *Gender-Related Bias in the Instrument: DIF Analysis*

The analysis of Differential Item Functioning (DIF Measure) revealed performance differences between male and female students in several physics test items. Items exhibiting substantial DIF include S14, S1, S6, and S2, which assess electromagnetic induction, magnetic fields, magnetic force between current-carrying wires, and the trajectory of charged particles in a magnetic field (The complete set of test items is available in the Appendix). These differences may be determined by factors such as mathematical proficiency, spatial reasoning skills, and previous experience with laboratory experiments.

On Item S14, which assesses Faraday's Law of Induction, male students demonstrated success in manipulating abstract mathematical concepts, particularly sinusoidal functions, whereas female students benefited more from visual representations and contextual explanations. The absence of illustrations in the questions may present a challenge for students in understanding how magnetic fields change over time. Similarly, Item S1, which evaluates the magnetic field pattern around a bar magnet, also exhibited gender-based performance disparities. The correct answer is "Location 4," where the magnetic field is strongest near the magnet's poles. The observed differences suggest that spatial reasoning and experimental exposure play an important role. Male students are generally more familiar with spatial representations, while female students often rely on narrative-based or hands-on experience approaches that may be less evident in the construction of test items.



A similar pattern was observed in Item S6, which assesses Ampere’s Law and the Lorentz Force. The correct answer, “the magnetic force decreases by one-fourth,” follows the inverse square law ( $1/r^2$ ). Male students demonstrated greater confidence in quadratic relationships and mathematical symbolism, whereas female students tended to perform better when the concepts were presented through direct experimentation or real-world applications. Without prior hands-on experience in observing the interaction of magnetic forces between two current-carrying wires, female students may have found it more difficult to reconcile the differences between theory and application.

Likewise, Item S2, which examines the trajectory of charged particles in a magnetic field based on the Lorentz force, revealed that female students faced greater difficulty in interpreting vector directions and visualizing three-dimensional motion. The correct response requires the application of the right-hand rule, which determines the trajectory based on the charge of the particle. Multirepresentational ability was a key challenge, with male students demonstrating greater familiarity with vector interpretation, while female students benefited more from interactive simulations and practical demonstrations to reinforce their conceptual understanding.

This finding aligns with research highlighting the impact of contextual familiarity and prior experiences on gender disparities in analytical assessments. Gender differences in physics self-efficacy have been linked to varied experiences (Nissen, 2019), while gender roles and biases influence engagement and performance (Musters et al., 2024). Similarly, contextual familiarity has been identified as a key contributor to gender gaps in physics education, reinforcing the importance of careful item review to ensure fair and equitable assessments (Taufik & Sutrio, 2023).

While most items exhibited minimal DIF, these findings emphasize the necessity of eliminating unintentional biases. Assessments should undergo systematic review to prevent embedded gender biases that could compromise fairness (Henderson et al., 2019; Widodo et al., 2024). Addressing such disparities is critical for ensuring that assessments are equitable across demographic groups. Gender biases in the Force Concept Inventory have been shown to significantly influence student performance (Mears, 2019), while social identity factors, including gender and race, play a pivotal role in assessment outcomes (Buncher et al., 2021). Additionally, eliminating hidden gender biases in science education is essential for fostering equity in learning environments (Nasri et al., 2023). To mitigate bias, revising items by removing context-specific elements or modifying language that may favor one gender is essential for creating a more inclusive assessment. Ensuring linguistic neutrality and contextual balance will enhance the instrument’s fairness and validity across diverse student populations.

### *Implications for Physics Assessment and Instruction*

This study establishes an integrated workflow that triangulates Bloom’s analytical framework, Rasch modeling, and Differential Item Functioning (DIF) diagnostics, thereby producing evidence-centered instruments that deliver defensible inferences about specific analytical sub-skills. In practice, such fine-grained feedback enables instructors to move assessment from a purely summative audit to a formative tool that pinpoints which students need remediation or enrichment. Because the calibrated items are already aligned with national competency standards, especially the emphasis on problem-solving and scientific reasoning, they can be embedded directly into magnetism units, lesson-study cycles, or digital learning platforms without requiring teachers to design new tasks from scratch. Finally, the systematic use of DIF screening throughout item development demonstrates a viable pathway toward equity-driven test construction, ensuring that gender or other demographic factors do not confound ability estimates and advancing current calls for bias-aware STEM evaluation.

### *Limitations and Future Directions*

Although the sample size ( $N = 84$ ) meets the stability requirements for Rasch calibration (Linacre, 1994), it was drawn from a single province, which may restrict generalizability. Methodologically, the present instrument relies exclusively on multiple-choice items; integrating constructed-response formats or interactive augmented-reality tasks could capture richer manifestations of analytical reasoning, such as causal chaining and model-based explanation. Future research might also explore multistage or fully adaptive algorithms that draw on the calibrated item bank to deliver personalized difficulty trajectories, thereby enhancing both measurement precision and instructional relevance. Collectively, these avenues offer promising routes for refining the instrument and extending its utility as a robust, equitable measure of higher-order thinking in physics education.

## Conclusions and Implications

This study successfully validated a Rasch-based analytical skills test for assessing higher-order thinking in the context of magnetism. The findings confirm that the instrument demonstrates high reliability (Cronbach's  $\alpha = .91$ , item reliability = .94) and effectively differentiates students' analytical abilities across various proficiency levels. The unidimensionality test and inter-item correlations further validate its construct alignment with the Bloom's Taxonomy dimensions of Differentiating, Organizing, and Attributing. However, Wright Map analysis identified gaps in item difficulty distribution, particularly in assessing extreme ability groups, highlighting the need for further refinement. Additionally, Differential Item Functioning (DIF) analysis detected gender biases in specific items, reinforcing the necessity of revisions to enhance fairness. These insights emphasize the significance of employing psychometrically rigorous assessments in physics education to ensure accurate measurement of analytical reasoning skills. Future research should prioritize optimizing item difficulty distribution, expanding the instrument to broader STEM domains, and exploring adaptive testing approaches to improve inclusivity and precision in student assessment.

This study suggests several refinements to enhance an instrument measuring analytical skills in physics. To improve inclusivity, it recommends developing items targeting a broader range of abilities, particularly at the extremes. Additionally, problematic items should be revised to better align with the Rasch model, and potential gender bias should be eliminated through strategies such as content-neutral scenarios and diverse pilot testing. Furthermore, expanding item coverage across all levels of Bloom's Taxonomy will provide a more comprehensive assessment of analytical skills. This study contributes to the growing body of research on Rasch-based assessments and suggests that future research explore adaptive testing and expand the instrument's scope to other physics topics or STEM fields.

## Acknowledgements

The researchers extend their sincere appreciation to the Indonesian Education Scholarship Program (BPI) under the Center for Higher Education Financing and Assessment (PPAPT), Secretariat General of the Ministry of Education, Culture, Research, and Technology, along with the Indonesia Endowment Fund for Education (LPDP), contract no: 03444/J5.2.3./BPI.06/10/2022 for their financial support in facilitating this research.

## Declaration of Interest

The authors declare no competing interest.

## References

- Abdellatif, M. S. M., & Abdel-Gawad, M. A. Z. (2020). The relative contribution of higher-order thinking in predicting psychological hardness of university students. *Asian Social Science*, 16(8), 68–77. <https://doi.org/10.5539/ass.v16n8p68>
- Agustihana, S., & Suparno, S. (2019). Development of HOTS oriented cognitive problems in thermodynamics for senior high schools. *Jurnal Penelitian Fisika Dan Aplikasinya (JPFA)*, 9(1), 44–54. <https://doi.org/10.26740/jpfa.v9n1.p44-54>
- Alanazi, A. A., Osman, K., & Halim, L. (2024). Effect of scaffolding strategies and guided discovery on higher-order thinking skills in physics education. *Eurasia Journal of Mathematics, Science and Technology Education*, 20(9), Article em2496. <https://doi.org/10.29333/ejmste/14980>
- Alhadabi, A., & Aldhafri, S. (2021). A Rasch model analysis of the psychometric properties of the Student-Teacher Relationship Scale among middle school students. *European Journal of Educational Research*, 10(2), 957–973. <https://doi.org/10.12973/eu-jer.10.2.957>
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Abridged ed.)*. Addison Wesley Longman. <https://eduq.info/xmlui/handle/11515/18345>
- Bond, T., & Fox, C. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3<sup>rd</sup> ed.). Routledge.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. <https://doi.org/10.1007/978-94-007-6857-4>
- Boone, W., Staver, J., & Yale, M. (2014). Person reliability, item reliability, and more (pp. 217–234). In *Rasch analysis in the human sciences*. Springer. [https://doi.org/10.1007/978-94-007-6857-4\\_10](https://doi.org/10.1007/978-94-007-6857-4_10)
- Buncher, J. B., Nissen, J. M., Van Dusen, B., Talbot, R. M. I., & Huvard, H. (2021). Bias on the Force Concept Inventory across the intersection of gender and race. *Physics Education Research Conference Proceedings*, 69–74. <https://www.per-central.org/items/detail.cfm?ID=15727>



- Can, D. (2020). The mediator effect of reading comprehension in the relationship between logical reasoning and word problem solving. *Participatory Educational Research*, 7(3), 230–246. <https://doi.org/10.17275/per.20.44.7.3>
- Chan, W. K., Zhang, L., & Oon, E. (2023). Classroom assessment that tailor instruction and direct learning: A validation study. *International Journal of Assessment Tools in Education*, 10(2), 376–394. <https://doi.org/10.21449/ijate.1155679>
- Chen, W., Zhang, G., Tian, X., Wang, L., & Luo, J. (2021). Rasch analysis of Work-Family Conflict Scale among Chinese prison police. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.537005>
- Churches, A. (2008). *Bloom's Digital Taxonomy*. <https://educationtechnologysolutions.com/2015/02/blooms-taxonomy-and-the-digital-world/>
- Damayanti, N., Hartono, Subali, B., Nugroho, S. E., & Sureeporn, K. (2020). Items analysis of physics assessment based on cognitive level of high order thinking skills in Bloom taxonomy. *Journal of Physics: Conference Series*, 1521(2), Article 022022. <https://doi.org/10.1088/1742-6596/1521/2/022022>
- Darmaji, D., Kurniawan, D. A., Astalini, A., Lumbantoruan, A., & Samosir, S. C. (2019). Mobile learning in higher education for the Industrial Revolution 4.0: Perception and response of physics practicum. *International Journal of Interactive Mobile Technologies (IJIM)*, 13(09), Article 09. <https://doi.org/10.3991/ijim.v13i09.10948>
- Denovan, A., Dagnall, N., & Drinkwater, K. (2022). The Ego Resiliency Scale-Revised: Confirmatory factor analysis and Rasch models. *Journal of Psychoeducational Assessment*, 40(6), 707–721. <https://doi.org/10.1177/07342829221090117>
- Denovan, A., Dagnall, N., Drinkwater, K., & Escolà-Gascón, Á. (2024). The Inventory of Personality Organization-Reality Testing Subscale and Belief in Science Scale: Confirmatory factor and Rasch analysis of thinking style measures. *PLOS ONE*, 19(9), Article e0310055. <https://doi.org/10.1371/journal.pone.0310055>
- Dessie, E., Gebeyehu, D., & Eshetu, F. (2023). Enhancing critical thinking, metacognition, and conceptual understanding in introductory physics: The impact of direct and experiential instructional models. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), Article em2287. <https://doi.org/10.29333/ejmste/13273>
- DeWitt, J., Archer, L., & Moote, J. (2019). 15/16-year-old students' reasons for choosing and not choosing physics at A level. *International Journal of Science and Mathematics Education*, 17(6), 1071–1087. <https://doi.org/10.1007/s10763-018-9900-4>
- Diniz, K. K. S., Pagano, A. S., Fernandes, A. P. P. C., Reis, I. A., Pinheiro Júnior, L. G., & Torres, H. de C. (2019). Development and validation of an instrument to assess Brazilian healthcare professional providers' knowledge on sickle cell disease. *Hematology, Transfusion and Cell Therapy*, 41(2), 145–152. <https://doi.org/10.1016/j.htct.2018.08.003>
- Erfianti, L., Istiyono, E., & Kuswanto, H. (2019). Developing LUP instrument test to measure higher order thinking skills (HOTS) Bloomian for senior high school students. *International Journal of Educational Research Review*, 4(3), 320–329. <https://doi.org/10.24331/ijere.573863>
- Forehand, M. (2010). Bloom's taxonomy. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology*. University of Georgia. [https://textbookequity.org/Textbooks/Orey\\_Emergin\\_Perspectives\\_Learning.pdf](https://textbookequity.org/Textbooks/Orey_Emergin_Perspectives_Learning.pdf)
- Ghanizadeh, A. (2017). The interplay between reflective thinking, critical thinking, self-monitoring, and academic achievement in higher education. *Higher Education*, 74(1), 101–114. <https://doi.org/10.1007/s10734-016-0031-y>
- Giancoli, D. C. (2014). *Physics: Principles with applications*. Pearson.
- Halliday, D., Resnick, R., & Walker, J. (2013). *Fundamentals of physics extended* (10th ed.). Wiley. [https://elearn.daffodilvarsity.edu.bd/pluginfile.php/987150/mod\\_label/intro/fundamentals-of-physics-textbook.pdf](https://elearn.daffodilvarsity.edu.bd/pluginfile.php/987150/mod_label/intro/fundamentals-of-physics-textbook.pdf)
- Henderson, R., Stewart, J., & Traxler, A. (2019). Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism. *Physical Review Physics Education Research*, 15(1), Article 010131. <https://doi.org/10.1103/PhysRevPhysEducRes.15.010131>
- Jamar, N. Z. Z., & Siew, N. M. (2020). Investigating the validity and reliability of a scientific imagination test for tenth graders. *International Journal of Educational Technology and Learning*, 8(2), 52–60. <https://doi.org/10.20448/2003.82.52.60>
- Kacprzyk, J., Parsons, M., Maguire, P. B., & Stewart, G. S. (2019). Examining gender effects in different types of undergraduate science assessment. *Irish Educational Studies*, 38(4), 467–480. <https://doi.org/10.1080/03323315.2019.1645721>
- Kamid, K., Kurniawan, D. A., & Nawahdani, A. M. (2022). Scientific learning model: Analytical thinking and process skills in mathematics. *Journal of Education Research and Evaluation*, 6(3), 538–549. <https://doi.org/10.23887/jere.v6i3.49159>
- Kliziene, I., Paskovske, A., Cizauskas, G., Augustiniene, A., Simonaitiene, B., & Kubiliunas, R. (2022). The impact of achievements in mathematics on cognitive ability in primary school. *Brain Sciences*, 12(6), Article 736. <https://doi.org/10.3390/brainsci12060736>
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. [https://doi.org/10.1207/s15430421tip4104\\_2](https://doi.org/10.1207/s15430421tip4104_2)
- Krell, M., Khan, S., & van Driel, J. (2021). Analyzing cognitive demands of a scientific reasoning test using the linear logistic test model (LLTM). *Education Sciences*, 11(9), Article 472. <https://doi.org/10.3390/educsci11090472>
- Kurniawan, C., Setyosari, P., Ulfa, S., Purnomo, Janan, D., Ridhoni, W., & Rahayu, S. (2022). Exploring the habit of using online learning platforms for higher order thinking skills with data mining techniques. In *2022 8th International Conference on Education and Technology (ICET)* (pp. 315–320). IEEE. <https://doi.org/10.1109/ICET56879.2022.9990654>
- Kwarikunda, D., Schiefele, U., Muwonge, C. M., & Ssenyonga, J. (2022). Profiles of learners based on their cognitive and metacognitive learning strategy use: Occurrence and relations with gender, intrinsic motivation, and perceived autonomy support. *Humanities and Social Sciences Communications*, 9(1), 1–12. <https://doi.org/10.1057/s41599-022-01322-1>
- Lestari, A., & Samsudin, A. (2020, July 30). Using Rasch model analysis to analyze students' scientific literacy on heat and temperature. In *Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar (MSCEIS 2019)*, 12 October 2019, Bandung, West Java, Indonesia. <https://eudl.eu/doi/10.4108/eai.12-10-2019.2296483>

- Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), Article 328. [https://www.researchgate.net/publication/285796448\\_Sample\\_size\\_and\\_item\\_calibration\\_or\\_person\\_measure\\_stability](https://www.researchgate.net/publication/285796448_Sample_size_and_item_calibration_or_person_measure_stability)
- Mancuso, M., Damora, A., Abbruzzese, L., Navarrete, E., Basagni, B., Galardi, G., Caputo, M., Bartalini, B., Bartolo, M., Zucchella, C., Carboncini, M. C., Dei, S., Zoccolotti, P., Antonucci, G., & De Tanti, A. (2019). A new standardization of the Bells Test: An Italian multi-center normative study. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.02745>
- Maulana, I., Siswandari, Gunarhadi, & Efendi, A. (2025). Virtual–Augmented Reality (VAR) for science learning: Development and impact on students' HOTS skills. *Electronic Journal of E-Learning*, 23(1), 129–142. <https://doi.org/10.34190/ejel.23.1.3733>
- Mears, M. (2019). Gender differences in the Force Concept Inventory for different educational levels in the United Kingdom. *Physical Review Physics Education Research*, 15(2), Article 020135. <https://doi.org/10.1103/PhysRevPhysEduRes.15.020135>
- Moussa, N. (2021). Assessing the higher education settings after the transition to online learning: Exploring teaching, assessments, and students' academic success. *International Journal of Learning, Teaching and Educational Research*, 20, 159–173. <https://doi.org/10.26803/ijlter.20.12.10>
- Musters, N., Aarts, R., Van Amelsvoort, M., & Swerts, M. (2024). Views on gender differences in the physics classroom. *Education Sciences*, 14(5), Article 457. <https://doi.org/10.3390/educsci14050457>
- Nasri, N., Rahimi, N. M., Hashim, H., & Nasri, N. M. (2023). "Supporting our lost boys": A research on gender-based science education's hidden curriculum in Malaysia. *International Journal of Learning, Teaching and Educational Research*, 22(3), 160–174. <https://www.ijlter.org/index.php/ijlter/article/view/6776>
- Nissen, J. M. (2019). Gender differences in self-efficacy states in high school physics. *Physical Review Physics Education Research*, 15(1), Article 013102. <https://doi.org/10.1103/PhysRevPhysEduRes.15.013102>
- Opfermann, M., Schmeck, A., & Fischer, H. E. (2017). Multiple representations in physics and science education – Why should we use them? In D. F. Treagust, R. Duit, & H. E. Fischer (Eds.), *Multiple representations in physics education* (pp. 1–22). Springer International Publishing. [https://doi.org/10.1007/978-3-319-58914-5\\_1](https://doi.org/10.1007/978-3-319-58914-5_1)
- Oreondo, L. L., & Antonio, E. M. D. (1984). *Evaluating educational outcomes*. Rex Book Store. <https://books.google.co.id/books?id=8xNMBn4bn8oC>
- Pullicino, N., & Bonello, C. (2020). Challenges faced by Maltese students studying advanced level physics. *Information*, 11(8), Article 397. <https://doi.org/10.3390/info11080397>
- Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research*, 8(3), 743–751. <https://doi.org/10.12973/eu-jer.8.3.743>
- Ridhoni, W., Setyosari, P., Kuswandi, D., Ulfa, S., & Janan, D. (2023). The analysis of test items for student's creative and critical thinking assessment: Towards mobile application development. *AIP Conference Proceedings*, 2679(1), Article 060012. <https://doi.org/10.1063/5.0111298>
- Sari, S. Y., Rahim, F. R., Sundari, P. D., & Aulia, F. (2022). The importance of e-books in improving students' skills in physics learning in the 21st century: A literature review. *Journal of Physics: Conference Series*, 2309(1), Article 012061. <https://doi.org/10.1088/1742-6596/2309/1/012061>
- Schwippert, K., Zilz, K., & Hoettecke, D. (2024). Difficulty-generating features of text-based physics tasks. *Journal of Research in Science, Mathematics and Technology Education*, 7(2), 1–24. <https://doi.org/10.31756/jrsmte.721>
- Solé-Llussà, A., Aguilar, D., & Ibáñez, M. (2022). Video-worked examples to support the development of elementary students' science process skills: A case study in an inquiry activity on electrical circuits. *Research in Science & Technological Education*, 40(2), 251–271. <https://doi.org/10.1080/02635143.2020.1786361>
- Spencer, M., Fuchs, L. S., & Fuchs, D. (2020). Language-related longitudinal predictors of arithmetic word problem solving: A structural equation modeling approach. *Contemporary Educational Psychology*, 60, Article 101825. <https://doi.org/10.1016/j.cedpsych.2019.101825>
- Surahman, E., Wedi, A., Sulthoni, Soepriyanto, Y., & Setyosari, P. (2018). Design of peer collaborative authentic assessment model based on group project-based learning to train higher order thinking skills of students. In *Proceedings of the International Conference on Education and Technology (ICET 2018)* (pp. 75–78). Atlantis Press. <https://www.atlantispress.com/proceedings/icet-18/125926623>
- Susiloningsih, W., Susilo, H., Zainuddin, M., Kuswandi, D., Suciptaningsih, O. A., & Faizah, H. (2024). The effect of cognitive style on the analysis ability of grade 4 elementary school students. *JPPi (Jurnal Penelitian Pendidikan Indonesia)*, 10(2), 490–495. <https://doi.org/10.29210/020243727>
- Taufik, M., & Sutrio, S. (2023). Deconstructing gender gaps in basic physics education: A multivariate analysis of practical work performance, semester outcomes and literacy among undergraduate biology education students. *Jurnal Ilmiah Profesi Pendidikan*, 8(3), 1976–1981. <https://doi.org/10.29303/jipp.v8i3.1117>
- Tavares-Júnior, J. W. L., de Souza, A. C. C., Alves, G. S., Bonfadini, J. de C., Siqueira-Neto, J. I., & Braga-Neto, P. (2019). Cognitive assessment tools for screening older adults with low levels of education: A critical review. *Frontiers in Psychiatry*, 10. <https://doi.org/10.3389/fpsyg.2019.00878>
- Tennant, A., & Küçükdeveci, A. A. (2023). A Application of the Rasch measurement model in rehabilitation research and practice: Early developments, current practice, and future challenges. *Frontiers in Rehabilitation Sciences*, 4. <https://doi.org/10.3389/fresc.2023.1208670>
- Tientongdee, S. (2018). Development of problem-solving skill by using active learning for student teachers in Introductory Physics. *Journal of Physics: Conference Series*, 1144(1), Article 012002. <https://doi.org/10.1088/1742-6596/1144/1/012002>
- Wheeler, S. R., & Blanchard, M. R. (2019). Contextual choices in online physics problems: Promising insights into closing the gender gap. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.00594>



Widodo, W., Sudana Degeng, I. N., Pali, M., & Hitipeuw, I. (2024). Development and validation of the Student Well-Being Scale for elementary school using Rasch model analysis. *Jurnal Educatio: Jurnal Edukasi dan Teknologi Pembelajaran*, 13(2). <https://doi.org/10.12928/jehcp.v13i2.28768>

Yuhanna, W. L., Muhdhar, M. H. I. A., Gofur, A., & Hassan, Z. (2021). S Self-reflection assessment in vertebrate zoology (SRAVZ) using Rasch analysis. *Jurnal Pendidikan IPA Indonesia*, 10(1), 35–47. <https://doi.org/10.15294/jpii.v10i1.25603>

Received: February 22, 2025

Revised: April 03, 2025

Accepted: May 11, 2025

Cite as: Nurussaniah, N., Setyosari, P., Kuswandi, D., & Ulfa, S. (2025). Psychometric validation of an analytical skills test in physics using the RASCH model. *Journal of Baltic Science Education*, 24(3), 522–537. <https://doi.org/10.33225/jbse/25.24.522>



**Nurussaniah Nurussaniah**  
(Corresponding author)

PhD Student, Department of Educational Technology, State University of Malang (Universitas Negeri Malang), Jl. Semarang 5 Malang 65145, East Java, Indonesia.

E-mail: nurussaniah.2201219@students.um.ac.id

ORCID: <https://orcid.org/0000-0003-2192-3691>

**Punaji Setyosari**

PhD, Professor, Department of Educational Technology, State University of Malang (Universitas Negeri Malang), Jl. Semarang 5 Malang 65145, East Java, Indonesia.

E-mail: punaji.setyosari.fip@um.ac.id

ORCID: <https://orcid.org/0000-0003-0187-9785>

**Dedi Kuswandi**

PhD, Professor, Department of Educational Technology, State University of Malang (Universitas Negeri Malang), Jl. Semarang 5 Malang 65145, East Java, Indonesia.

E-mail: dedi.kuswandi.fip@um.ac.id

ORCID: <https://orcid.org/0000-0003-1005-6641>

**Saida Ulfa**

PhD, Associate Professor, Department of Educational Technology, State University of Malang (Universitas Negeri Malang), Jl. Semarang 5 Malang 65145, East Java, Indonesia.

E-mail: saida.ulfa.fip@um.ac.id

ORCID: <https://orcid.org/0000-0002-2302-7172>

