

## Measuring Scientific Reasoning in Indonesian Senior High School Students: A Rasch Model Analysis

Risnatati<sup>1</sup>, Ridwan Efendi<sup>1\*</sup>, Winny Liliawati<sup>1</sup>, Achmad Samsudin<sup>1</sup>, and Duden Saepuzaman<sup>1</sup>

<sup>1</sup>Physics Education Study Program, Universitas Pendidikan Indonesia, Indonesia

\*[ridwanefendi@upi.edu](mailto:ridwanefendi@upi.edu)

Received  
28/05/2026

Accepted  
25/08/2026

Published  
10/10/2026

DOI  
[10.59329/gawi.v6i2.323](https://doi.org/10.59329/gawi.v6i2.323)

### ABSTRACT

This study aimed to evaluate the psychometric quality of a scientific reasoning instrument using the Rasch measurement model. A descriptive survey method with a non-experimental cross-sectional design was employed. The participants comprised 65 eleventh-grade senior high school students from a single school in Central Sulawesi, Indonesia, selected through purposive sampling. The instrument was adapted from Lawson's Classroom Test of Scientific Reasoning (LCTSR) and consisted of 12 multiple-choice items covering four dimensions of scientific reasoning: proportional reasoning, control of variables, combinatorial reasoning, and probabilistic reasoning. Data were analyzed using Rasch model analysis to examine reliability, item fit, person ability, and unidimensionality. The results indicated functional but imperfect measurement quality: person reliability was moderate to weak (0.63), item reliability was strong (0.92), and Cronbach's alpha (0.66) reflected acceptable internal consistency. One item (Q6) showed clear misfit (Outfit MNSQ = 2.39), and two items (Q4 and I0011) showed weaker-than-ideal discrimination (PTMEA CORR = 0.31), indicating that not all items met the fit criteria established in this study. Students' ability estimates ranged from -2.85 to +2.93 logits, reflecting heterogeneous levels of scientific reasoning ability, and the Wright Map showed a reasonably good, though imperfect, alignment between student ability and item difficulty. Item Q6, representing the control-of-variables dimension, was identified as the most difficult and most misfitting item, while item Q7, representing combinatorial reasoning, was the easiest, suggesting that students experienced greater difficulty in causal reasoning and variable control than in combinatorial reasoning tasks. Differential Item Functioning (DIF) analysis by gender indicated no statistically significant bias for most items, although item Q7 showed a significant DIF contrast on one of two tests used; this result should be interpreted cautiously, given the small and unbalanced gender subgroups. These findings suggest that scientific reasoning develops unevenly across dimensions and may be influenced by task complexity, while also highlighting specific items requiring further refinement before the instrument is used more broadly.

**Keywords:** *Lawson test, psychometric analysis, Rasch model, scientific reasoning, Wright Map*

### How to cite:

Risnatati, Efendi, R., Liliawati, W., Samsudin, A., & Saepuzaman, D. (2026). Measuring Scientific Reasoning in Indonesian Senior High School Students: A Rasch Model Analysis. *Gawi: Journal of Action Research*, 6 (2), 346-360



## INTRODUCTION

Among the various competencies emphasized in 21st-century education, scientific reasoning and critical thinking are considered essential skills that support problem solving, decision making, and innovation ([Facione, 1990](#)). Scientific reasoning (SR) refers to a set of cognitive processes used to construct scientific knowledge through evidence-based thinking and systematic investigation of natural phenomena ([Brown et al., 2010](#); [Díaz et al., 2023](#)). These skills involve evaluating evidence, formulating hypotheses, designing experiments, analyzing data, and drawing logical conclusions, all of which are fundamental for adapting to the challenges of the digital and globalized era ([Choowong & Worapun, 2021](#)). In science education, scientific reasoning is closely associated with students' ability to solve authentic problems and is an important predictor of academic achievement and long-term success in science learning ([Abate et al., 2020](#)).

The development of scientific reasoning skills has therefore become one of the primary goals of science education across different educational levels and countries. Previous studies have investigated various instructional approaches to improve students' scientific reasoning abilities ([Chen & She, 2015](#); [Sapia et al., 2022](#)). Research findings indicate that students with stronger scientific reasoning skills tend to demonstrate better performance in understanding scientific concepts and solving real-world problems ([Han, 2013](#); [Zhou et al., 2016](#)). However, improving scientific reasoning requires not only effective instruction but also valid and reliable assessment tools capable of accurately measuring students' reasoning abilities ([Abate et al., 2020](#)). Despite growing attention to scientific reasoning, assessing these skills remains challenging because it involves multiple interconnected cognitive processes that cannot be adequately captured by conventional assessments.

Conceptually, scientific reasoning can be viewed as part of mechanistic reasoning, particularly in understanding causal relationships within scientific phenomena ([Bachtiar et al., 2022](#)). Previous studies have suggested that scientific reasoning represents a universal cognitive process that supports evidence-based argumentation and scientific inquiry ([Omarchevska et al., 2022](#)). In this study, the scientific reasoning framework integrates the cause-and-effect approach with Lawson's framework. The cause-and-effect reasoning perspective explains how students understand relationships between variables and construct explanations based on evidence and underlying mechanisms ([Bao et al., 2022](#); [Giri & Paily, 2020](#)). Meanwhile, Lawson's framework, which is grounded in Piaget and Inhelder's cognitive development theory, categorizes scientific reasoning into several indicators, namely proportional reasoning, control of variables, combinatorial reasoning, and probabilistic reasoning ([Khan & Rana, 2021](#); [Xiao et al., 2018](#)). These indicators represent progressive levels of scientific reasoning development, ranging from basic to more complex reasoning processes.

The integration of these frameworks enables scientific reasoning to be analyzed not only from the perspective of measurement outcomes but also from the cognitive processes underlying students' responses. To assess students' scientific reasoning abilities, this study employed the Lawson Classroom Test of Scientific Reasoning (LCTSR), which has been widely used and validated in science education research ([Sapia et al., 2022](#)). The LCTSR is designed to evaluate students' reasoning patterns across multiple dimensions of scientific reasoning through structured multiple-choice items. The conceptual framework of scientific reasoning adopted in this study is presented in [Figure 1](#).

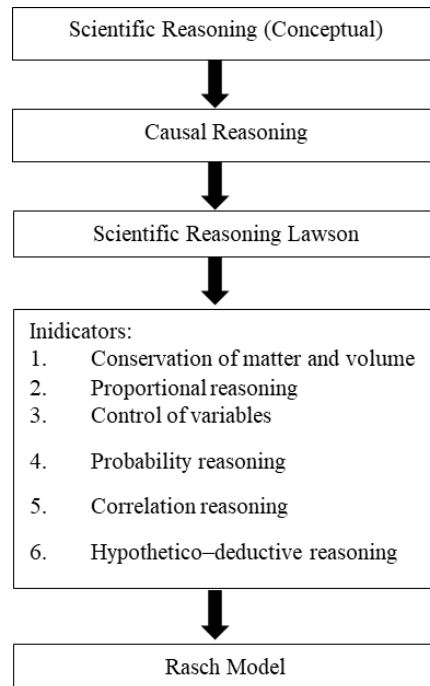


**Figure 1. Framework scientific reasoning**

Although several studies have examined scientific reasoning in Indonesia using conventional psychometric or descriptive scoring approaches, including studies employing the LCTSR itself among first-year university students ([Novia et al., 2018](#)). Limited studies have comprehensively evaluated the measurement quality of scientific reasoning instruments using the Rasch measurement model, particularly among secondary school students in the Indonesian context. Preliminary observations during science learning indicated that many students had difficulty with tasks involving variable control, probabilistic reasoning, and causal explanation. These observations suggest the need for a measurement approach that can simultaneously identify students' reasoning abilities and the psychometric quality of assessment items. Rasch analysis provides a powerful probabilistic measurement model for evaluating the interaction between person ability and item difficulty on a common logit scale ([Nurhayati et al., 2025](#); [Planinic et al., 2019](#)). Unlike classical test theory, the Rasch model allows researchers to evaluate item fit, person fit, unidimensionality, reliability, item difficulty hierarchy, and Differential Item Functioning (DIF) in a more objective and comprehensive manner. Through Wright Map analysis, the Rasch model also enables visualization of the relationship between students' abilities and item difficulty levels, thereby providing deeper insight into students' scientific reasoning patterns ([Boone et al., 2014b](#)).

Several studies in physics and science education have applied Rasch analysis in instrument development and validation, including scientific reasoning assessments outside the Indonesian context ([Abate et al., 2020](#); [Xiao et al., 2018](#)), alternative conception instruments in Indonesia ([Aminudin et al., 2019](#)), and conceptual change studies in augmented reality learning environments ([Samsudin et al., 2025](#)). However, to the best of the authors' knowledge, no published study has applied Rasch model analysis to evaluate the psychometric properties of the LCTSR specifically among Indonesian senior high school students a gap that is notable given that the instrument has previously been used in Indonesia only with university-level samples and analyzed using classical, descriptive approaches ([Novia et al., 2018](#)). This study, therefore, aims to evaluate the quality of a scientific reasoning instrument adapted from the LCTSR using Rasch model analysis, examining item fit, person fit, reliability, unidimensionality, Differential Item Functioning (DIF), and the distribution of item difficulty and student ability through Wright Map analysis. Beyond establishing the instrument's overall validity and reliability, this study further

examines the hierarchy of reasoning difficulty across Lawson's scientific reasoning dimensions through Wright Map interpretation, offering insight into students' cognitive reasoning patterns and the complexity of scientific reasoning indicators within the Indonesian secondary education context.

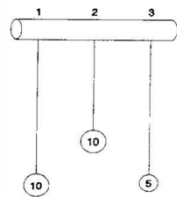
## METHOD

This study employed a descriptive survey method with a non-experimental cross-sectional design to evaluate senior high school students' scientific reasoning abilities in Central Sulawesi, Indonesia. Data were collected using a scientific reasoning instrument administered through Google Forms. The collected responses were analyzed using the Rasch measurement model to evaluate students' scientific reasoning abilities and the psychometric quality of the instrument. The instrument used in this study was the LCTSR, adapted from Lawson's scientific reasoning framework. The items were directly translated from English into Indonesian by the research team. Content and language validity were established through expert judgment involving two reviewers with expertise in science/physics education, who evaluated the translated items to ensure conceptual accuracy and to identify any potential translation errors prior to use. No separate readability testing or pilot administration was conducted following the expert review; the instrument was administered to the study sample immediately after the expert judgment process. The instrument consisted of 12 multiple-choice items covering four dimensions of scientific reasoning: proportional reasoning, control of variables, combinatorial reasoning, and probabilistic reasoning. The distribution of indicators and item numbers is presented in [Table 1](#) and an example test item is shown in [Figure 2](#).

**Table 1. Indicator Scientific Reasoning Skill**

| Dimention              | Item Number | Number of Item |
|------------------------|-------------|----------------|
| Proportional Reasoning | 1, 2, 3, 4  | 4              |
| Control of Variabel    | 5, 6        | 2              |
| Combination Reasoning  | 7, 8        | 2              |
| Probability Reasoning  | 9,10,11,12  | 4              |

Gambar di bawah menyajikan tiga senar yang tergantung pada sebuah balok. ketiga senar tersebut memiliki pemberat logam yang melekat pada ujungnya. Senar 1 dan senar 3 sama panjangnya. Senar 2 lebih pendek. Pemberat 10 gram menempel pada ujung senar 1. Pemberat 10 gram juga menempel pada ujung senar 2. Pemberat 5 gram menempel pada ujung senar 3. Senar (dan pembera yang terpasang) bisa diayun maju mundur dan waktu yang dibutuhkan senar tersebut untuk berayun dapat dihitung waktunya.



Misalkan Anda ingin mengetahui apakah Panjang tali memiliki pengaruh pada waktu yang dibutuhkan untuk Berayun maju mundur. Senar mana akan Anda gunakan untuk mencari tahu?

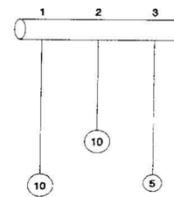
- hanya satu senar
- ketiga senar
- senar 2 dan 3
- senar 1 dan 3
- senar 1 dan 2

*karena*

- Anda harus menggunakan senar terpanjang.
- Anda harus membandingkan senar dengan bobot ringan dan berat.
- hanya panjangnya yang berbeda.
- untuk membuat semua kemungkinan perbandingan.
- bobotnya berbeda.

**a**

The figure below shows three strings hanging from a block. All three strings have metal weights attached to their ends. String 1 and string 3 are the same length. String 2 is shorter. A 10-gram weight is attached to the end of String 1. A 10-gram weight is also attached to the end of String 2. A 5-gram weight is attached to the end of String 3. The strings (and their attached weights) can be swung back and forth, and the time it takes for the strings to swing can be measured.



Suppose you want to find out whether the length of the string affects the time it takes to swing back and forth. Which strings would you use to find out?

- only one string
- all three strings
- strings 2 and 3
- strings 1 and 3
- strings 1 and 2

*because*

- You should use the longest string.
- You should compare the light and heavy strings.
- They differ only in length.
- To make all possible comparisons.
- They have different weights.

**b**

**Figure 2. Example of the Scientific Reasoning Instrument (a) Indonesian Version (b) English Version**

The participants of this study were 65 eleventh-grade (XI) senior high school students from one senior high school in Central Sulawesi, Indonesia. Participants were selected using purposive sampling, with the inclusion criteria being students from classes whose teachers consented to

participate and who were present during the test administration. Permission was obtained from the school prior to data collection, and all students provided informed consent before participating. Student responses were anonymized to protect participant confidentiality. Formal ethical clearance from an institutional ethics committee was not obtained for this study; ethical considerations were addressed through school permission and student informed consent.

Data analysis was conducted using the Rasch measurement model with Ministep software. Students' responses were scored dichotomously, with correct answers coded as 1 and incorrect answers coded as 0. The Rasch model was used to examine item difficulty, person ability, item fit, person fit, reliability, unidimensionality, and DIF. Instrument reliability was evaluated using person reliability, item reliability, and Cronbach's alpha coefficient. According to Sumintono and Widhiarso (2015), reliability values were categorized as weak ( $x < 0.67$ ), fair ( $0.67 \leq x \leq 0.80$ ), good ( $0.81 \leq x \leq 0.90$ ), very good ( $0.91 \leq x \leq 0.94$ ), and excellent ( $x > 0.94$ ). Cronbach's alpha values were interpreted based on Boone et al. (2014a), where  $\alpha < 0.50$  indicates low reliability,  $0.50 \leq \alpha < 0.60$  indicates moderate reliability,  $0.60 \leq \alpha < 0.70$  indicates acceptable reliability, and  $\alpha \geq 0.80$  indicates high reliability.

The fit of the data to the Rasch model was evaluated using Infit and Outfit Mean Square (MNSQ) statistics and standardized Z values (ZSTD). The data were considered fit to the model if the MNSQ values ranged between 0.5 and 1.5 and the ZSTD values ranged between  $-2$  and  $+2$  (Linacre, 1999). Unidimensionality was assessed using Principal Component Analysis of Residuals (PCAR), with criteria of raw variance explained by measures  $> 20\%$  and first contrast  $< 15\%$  (eigenvalue  $< 2.0$ ), indicating that the instrument measured a single latent construct. Item difficulty was determined based on item measure values expressed in logits, where positive logit values indicated more difficult items and negative logit values indicated easier items. Wright Map analysis was used to examine the alignment between item difficulty and students' ability levels. Item fit evaluation included Outfit MNSQ, Outfit ZSTD, and Point Measure Correlation (PTMEA CORR) values. Items were considered acceptable if they met the criteria of  $0.5 < \text{MNSQ} < 1.5$ ,  $-2 < \text{ZSTD} < +2$ , and  $0.4 < \text{PTMEA CORR} < 0.85$  (Amiruddin et al., 2023; Sumintono, 2018). Furthermore, DIF analysis based on gender was conducted to identify potential item bias between male and female students. Items were considered free from bias if the DIF contrast value was below 0.5 logits and the probability value exceeded 0.05 (Sumintono, 2018).

## RESULT AND DISCUSSION

### Summary Statistics

This study aimed to evaluate students' scientific reasoning abilities and examine the psychometric quality of the scientific reasoning instrument using the Rasch measurement model. The summary statistics of the Rasch analysis are presented in Table 2.

**Table 2. Summary statistics of person and items**

| Psychometrics Attribute | Person | Item  |
|-------------------------|--------|-------|
| Number                  | 65     | 12    |
| Measure (Logit) Mean    | -0.27  | 0.00  |
| SD, Standard Deviasi    | 1.30   | 1.14  |
| SE, Standard Error      | 0.74   | 0.31  |
| Mean INFIT MNSQ         | 0.98   | 1.01  |
| Mean INFIT ZSTD         | -0.02  | -0.13 |
| Mean OUTFIT MNSQ        | 1.13   | 1.13  |
| Mean OUTFIT ZSTD        | 0.10   | 0.12  |
| Separation              | 1.30   | 3.31  |
| Reliability             | 0.63   | 0.92  |
| Cronbach Alpha          | 0.66   |       |

The average INFIT and OUTFIT MNSQ values for both persons and items were close to 1.00, indicating that the response patterns were broadly consistent with model expectations. The item difficulty mean was centered at 0.00 logits with a standard deviation of 1.14, reflecting a fairly

wide spread of item difficulty levels consistent with the observed range of -2.07 to 2.61 logits. In terms of student performance, the average person measure was -0.27 logits, indicating that students' scientific reasoning abilities were, on average, slightly below the overall item difficulty level. The person reliability value of 0.63 indicates only moderate, rather than strongly acceptable, separation of ability levels; following Sumintono and Widhiarso (2015), categorization, this value falls in the "weak" range ( $\alpha < 0.67$ ). The separation index of 1.30 suggests the instrument distinguished students into approximately two ability strata, which is a modest level of discrimination for a 12-item instrument. This limited discrimination should be acknowledged as a constraint on the instrument's diagnostic precision at the individual level, rather than described as generally "acceptable." The Cronbach's alpha (0.66) falls in the "acceptable" band per Boone et al. (2014a), but this should be read alongside the moderate person reliability rather than in isolation. Thus, although the instrument demonstrated adequate overall internal consistency, its ability to consistently differentiate students according to their scientific reasoning ability remained limited.

### Unidimensionality

The construct validity was evaluated through unidimensionality analysis using Principal Component Analysis of Residuals (PCAR). The results are presented in Figure 3.

| STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units |            |          |          |
|-----------------------------------------------------------------------------|------------|----------|----------|
|                                                                             | Eigenvalue | Observed | Expected |
| Total raw variance in observations =                                        | 18.3000    | 100.0%   | 100.0%   |
| Raw variance explained by measures =                                        | 6.3000     | 34.4%    | 33.4%    |
| Raw variance explained by persons =                                         | 3.4663     | 18.9%    | 18.4%    |
| Raw Variance explained by items =                                           | 2.8336     | 15.5%    | 15.0%    |
| Raw unexplained variance (total) =                                          | 12.0000    | 65.6%    | 66.6%    |
| Unexplained variance in 1st contrast =                                      | 2.1159     | 11.6%    | 17.6%    |
| Unexplained variance in 2nd contrast =                                      | 1.8241     | 10.0%    | 15.2%    |
| Unexplained variance in 3rd contrast =                                      | 1.5355     | 8.4%     | 12.8%    |
| Unexplained variance in 4th contrast =                                      | 1.3741     | 7.5%     | 11.5%    |
| Unexplained variance in 5th contrast =                                      | 1.1325     | 6.2%     | 9.4%     |

Figure 3. Unidimensionality analysis

The raw variance explained by measures reached 34.4%, exceeding the minimum recommended criterion of 20% (Boone et al., 2014b), and the unexplained variance in the first contrast (11.6%) remained below the 15% threshold. However, the eigenvalue of the first contrast was 2.12, marginally exceeding the 2.0 threshold adopted in this study as a complementary criterion for unidimensionality. This indicates that while the percentage-based criterion was satisfied, the eigenvalue-based criterion was not fully met, suggesting the presence of a secondary dimension that warrants caution in interpretation. The essential unidimensionality index of 47.9% indicates that less than half of the total variance was explained by the primary construct measured by the Rasch model. This finding suggests that the instrument demonstrated a moderate, rather than strong, level of unidimensionality.

An inspection of standardized residual correlations identified a notable local item dependence between Q4 and Q6 (residual correlation = 0.47), exceeding the conventional flag threshold of 0.3 (Linacre, 2011). Given that LCTSR items are known to sometimes share conceptual or contextual overlap, this finding is reported as a limitation of the instrument's strict unidimensionality assumption; future studies should consider testlet modeling or further qualitative review of item content for Q4 and Q6. Although the instrument consisted of several dimensions of scientific reasoning, the empirical results suggest that these dimensions largely though not perfectly represent a shared underlying construct. This finding offers qualified support for the assumption that scientific reasoning functions as an integrated cognitive process (Kambeyo & Csapó, 2018;

[Koslowski, 1996](#)), while also indicating that some multidimensionality and item dependence remain and should be addressed in future refinement of the instrument.

### Item difficulty

The Rasch model also provided information regarding item difficulty and item fit. The results of the item analysis are presented in [Table 3](#).

**Table 3. Item difficulty and item fit analysis**

| No | Item  | Logit | SEM  | Outfit |       | PTMEA CORR |
|----|-------|-------|------|--------|-------|------------|
|    |       |       |      | MNSQ   | ZSTD  |            |
| 1  | Q6    | 2.61  | 0.46 | 2.39   | 1.56  | 0.09       |
| 2  | Q4    | 0.98  | 0.31 | 1.29   | 0.94  | 0.31       |
| 3  | Q1    | 0.70  | 0.30 | 1.32   | 1.15  | 0.31       |
| 4  | I0011 | 0.43  | 0.29 | 1.06   | 1.15  | 0.55       |
| 5  | I0012 | 0.43  | 0.29 | 1.17   | 0.76  | 0.46       |
| 6  | Q2    | 0.18  | 0.29 | 0.84   | -0.70 | 0.55       |
| 7  | Q3    | -0.22 | 0.28 | 1.84   | -0.77 | 0.53       |
| 8  | Q9    | -0.38 | 0.28 | 0.74   | -1.33 | 0.59       |
| 9  | Q8    | -0.55 | 0.29 | 0.95   | -0.17 | 0.51       |
| 10 | Q1    | -0.97 | 0.29 | 1.01   | 0.11  | 0.46       |
| 11 | Q5    | -1.14 | 0.30 | 0.76   | -0.90 | 0.57       |
| 12 | Q7    | -2.07 | 0.35 | 1.15   | 0.47  | 0.41       |

Applying the fit criteria established in this study ( $0.5 < \text{Outfit MNSQ} < 1.5$ ;  $-2 < \text{ZSTD} < +2$ ;  $0.4 < \text{PTMEA CORR} < 0.85$ ), the results show that item Q6 is the only item exceeding the Outfit MNSQ threshold ( $2.39 > 1.5$ ), confirming substantial misfit. In addition, three items fall below the PTMEA CORR threshold of 0.4: Q6 (0.09), Q4 (0.31), and I0011 (0.31). These items showed weaker-than-expected discrimination and are reported explicitly here rather than treated as part of a generally well-functioning item set.

The item difficulty values were distributed from  $-2.07$  to  $2.61$  logits, indicating that the instrument covered a fairly wide range of difficulty levels from easy to very difficult items. Item Q6 was identified as the most difficult item ( $2.61$  logits), while item Q7 was the easiest item ( $-2.07$  logits). Item Q6, which represented the control-of-variables dimension, demonstrated the highest difficulty level and also showed clear misfit, indicated by a high Outfit MNSQ value ( $2.39$ ) and a low PTMEA CORR value ( $0.09$ ) both outside the fit criteria adopted in this study. This finding suggests that students experienced substantial difficulties in reasoning tasks involving the identification and control of interacting variables within complex causal situations, although the item's misfit also means its difficulty estimate should be interpreted with some caution rather than taken at face value. From a cognitive perspective, control-of-variables reasoning requires students to coordinate multiple variables simultaneously and isolate causal relationships systematically, which are characteristics of higher-order formal operational thinking ([Schauble et al., 1991](#)). This finding is broadly consistent with previous studies reporting that variable control is among the most cognitively demanding dimensions of scientific reasoning ([Van Vo & Csapó, 2021](#); [Zhou et al., 2016](#)).

In contrast, item Q7 demonstrated the lowest difficulty level, indicating that students tended to perform better on combinatorial reasoning tasks involving pattern recognition and simple enumeration strategies. Not all items, however, met every fit criterion adopted in this study: alongside Q6, items Q4 and I0011 showed PTMEA CORR values ( $0.31$  each) below the  $0.4$  threshold, indicating comparatively weaker discrimination than the remaining items, even though their Outfit MNSQ values remained within range. The instrument's difficulty hierarchy is therefore reasonably broad and largely functional, but it is not uniformly well-behaved across all items, and the items identified above should be prioritized for qualitative review and, in the case of Q6, explicit caution in interpretation in any future revision of the instrument

### Differential Item Functioning (DIF)

Differential Item Functioning (DIF) analysis was conducted to identify potential item bias across gender groups. The probability values obtained from the DIF analysis are presented in [Figure 4](#).

| Person<br>CLASSES | SUMMARY DIF |      |       | BETWEEN-CLASS/GROUP |        |       | Item   |       |
|-------------------|-------------|------|-------|---------------------|--------|-------|--------|-------|
|                   | CHI-SQUARED | D.F. | PROB. | UNWTD               | MNSQ   | ZSTD  | Number | Name  |
| 2                 | .2592       | 1    | .6107 |                     | .2706  | -.28  | 1      | Q1    |
| 2                 | .0011       | 1    | .9735 |                     | .0013  | -1.42 | 2      | Q2    |
| 2                 | 1.3129      | 1    | .2519 |                     | 1.4240 | .74   | 3      | Q3    |
| 2                 | .0031       | 1    | .9553 |                     | .0056  | -1.27 | 4      | Q4    |
| 2                 | 1.4194      | 1    | .2335 |                     | 1.5445 | .80   | 5      | Q5    |
| 2                 | .0181       | 1    | .8930 |                     | .0188  | -1.09 | 6      | Q6    |
| 2                 | 4.2270      | 1    | .0398 |                     | 5.1320 | 2.01  | 7      | Q7    |
| 2                 | 1.4639      | 1    | .2263 |                     | 1.5994 | .83   | 8      | Q8    |
| 2                 | 2.8868      | 1    | .0893 |                     | 3.3076 | 1.51  | 9      | Q9    |
| 2                 | .1978       | 1    | .6565 |                     | .2062  | -.40  | 10     | Q10   |
| 2                 | .8892       | 1    | .3457 |                     | .9447  | .43   | 11     | I0011 |
| 2                 | .1978       | 1    | .6565 |                     | .2062  | -.40  | 12     | I0012 |

Figure 4. DIF probability values

The DIF analysis showed that all items had probability values greater than 0.05, indicating that no statistically significant DIF was detected between male and female students. These findings suggest that the instrument functioned fairly across gender groups and that the items measured scientific reasoning consistently regardless of gender differences. Most items showed in [Figure 5](#) only minor variations in difficulty between male and female students, and these differences remained within acceptable limits.

Person DIF plot

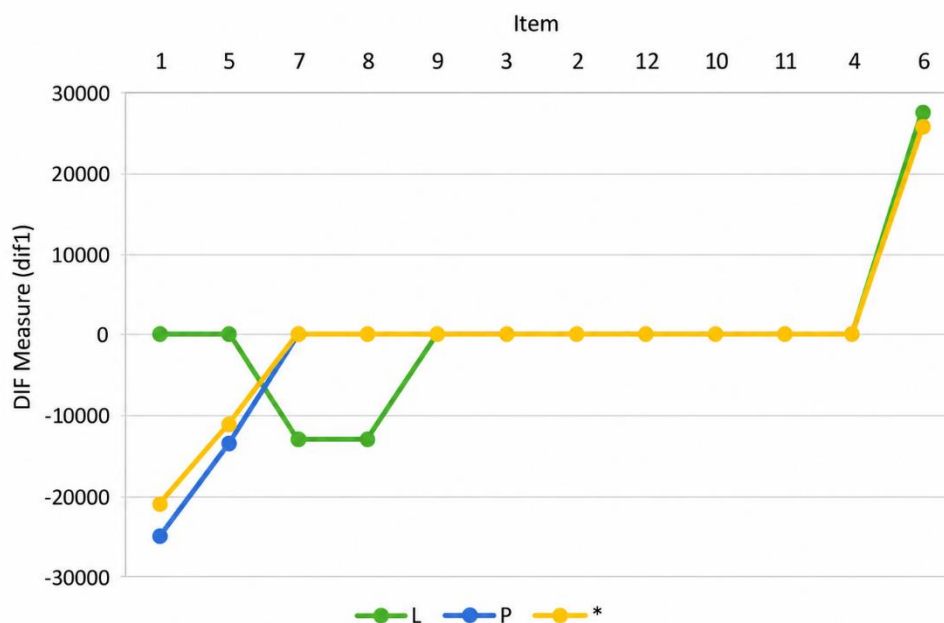


Figure 5. DIF analysis by gender

Item Q7, however, stands out: as a combinatorial reasoning item, its borderline-significant DIF result suggests that its wording or contextual framing may warrant qualitative review to determine whether it differentially affects male and female respondents. Given the small and unbalanced

subgroup sizes ( $n = 14$ ;  $n = 51$ ), this finding should be treated as exploratory rather than conclusive. Therefore, rather than describing the instrument as unconditionally fair and stable across gender groups, a more accurate conclusion is that the instrument performed consistently for most items, with one item (Q7) flagged for further review.

### Person Fit

Person fit analysis was conducted to evaluate whether students' response patterns were consistent with the expectations of the Rasch model. This analysis is important for identifying unusual response patterns that may indicate guessing behavior, carelessness, or inconsistent reasoning processes. The summary of the person fit analysis is presented in [Table 4](#).

**Tabel 4. Person Fit**

| Indicator          | Score | Criteria    | Description                        |
|--------------------|-------|-------------|------------------------------------|
| Responden          | 65    | -           | All data                           |
| Person reliability | 0,63  | $\geq 0,60$ | Weak to Moderate                   |
| Person separation  | 1,30  | $\geq 1,00$ | Distinguishing 2 levels of ability |
| Mean INFIT MNSQ    | 0,98  | 0,5 – 1,5   | Fit to model                       |
| Mean OUTFIT MNSQ   | 1,13  | 0,5 – 1,5   | Fit to model                       |

The average INFIT MNSQ value of 0.98 and OUTFIT MNSQ value of 1.13 indicate that, on average, students provided responses broadly consistent with the scientific reasoning construct measured by the instrument. However, individual inspection of person fit statistics identified several misfitting response patterns (e.g., persons 20, 9, 29, 37, with OUTFIT MNSQ values ranging from 1.42 to 9.90), suggesting that a subset of students exhibited response patterns inconsistent with the Rasch model, possibly reflecting guessing behavior or inconsistent reasoning strategies. The distribution of students' scientific reasoning abilities is presented in [Figure 6](#).

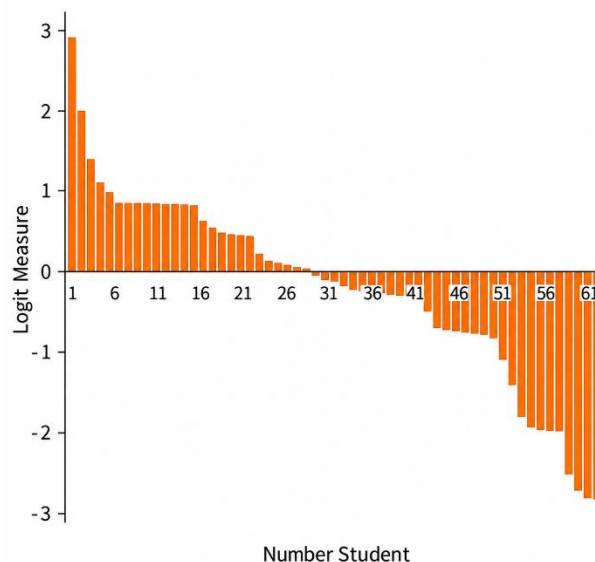


**Figure 6. Distribution of Students' Scientific Reasoning Ability**

The person measure distribution indicates substantial variation in students' scientific reasoning abilities, ranging from low to high ability levels. While the broad distribution of person logits suggests the instrument captured differences in reasoning performance, the moderate person reliability (0.63) and the presence of several misfitting persons indicate that some of the observed

variability may also reflect measurement imprecision rather than purely substantive differences in ability.

### Scalogram

The response patterns identified through the Guttman scalogram analysis are presented in [Figure 7](#).

GUTTMAN SCALOGRAM OF RESPONSES:

| Person | Item           |              |
|--------|----------------|--------------|
|        |                | 111          |
|        |                | 751893202146 |
|        |                | -----        |
| 7      | +1111111111110 | 07P          |
| 23     | +1111111111110 | 23P          |
| 49     | +1111111111100 | 49P          |
| 53     | +1111111101110 | 53P          |
| 17     | +1111111111000 | 17L          |
| 29     | +0111111111010 | 29P          |
| 37     | +1100110111111 | 37P          |
| 1      | +1011011001111 | 01L          |
| 4      | +1111111110000 | 04L          |
| 8      | +1111011110010 | 08P          |
| 13     | +1110111101001 | 13P          |
| 22     | +1111011101010 | 22P          |
| 38     | +1111101111000 | 38P          |
| 41     | +1111111100100 | 41P          |
| 43     | +1111111110000 | 43P          |
| 50     | +1111111011000 | 50L          |
| 60     | +1111111110000 | 60P          |
| 61     | +1111101111000 | 61P          |

**Figure 7. Scalogram of student responses**

The Guttman scalogram analysis showed that most students demonstrated response patterns that generally followed the expected hierarchy of item difficulty. Students with higher scientific reasoning ability tended to answer items sequentially from easier to more difficult items, while students with lower ability levels began to show incorrect responses on more complex items. This graded response pattern supports the unidimensionality of the instrument and indicates that the items consistently measured the same underlying construct ([Depaoli et al., 2018](#)). However, several inconsistent response patterns were still identified, suggesting the possibility of guessing behavior, misconceptions, or difficulties associated with specific scientific reasoning indicators. These findings indicate that non-cognitive factors may also influence students' responses and should be considered in future instrument refinement and instructional design.

### Wright Map

Figure 8 presents the Wright Map illustrating the relationship between students' ability levels and item difficulty on a common logit scale.

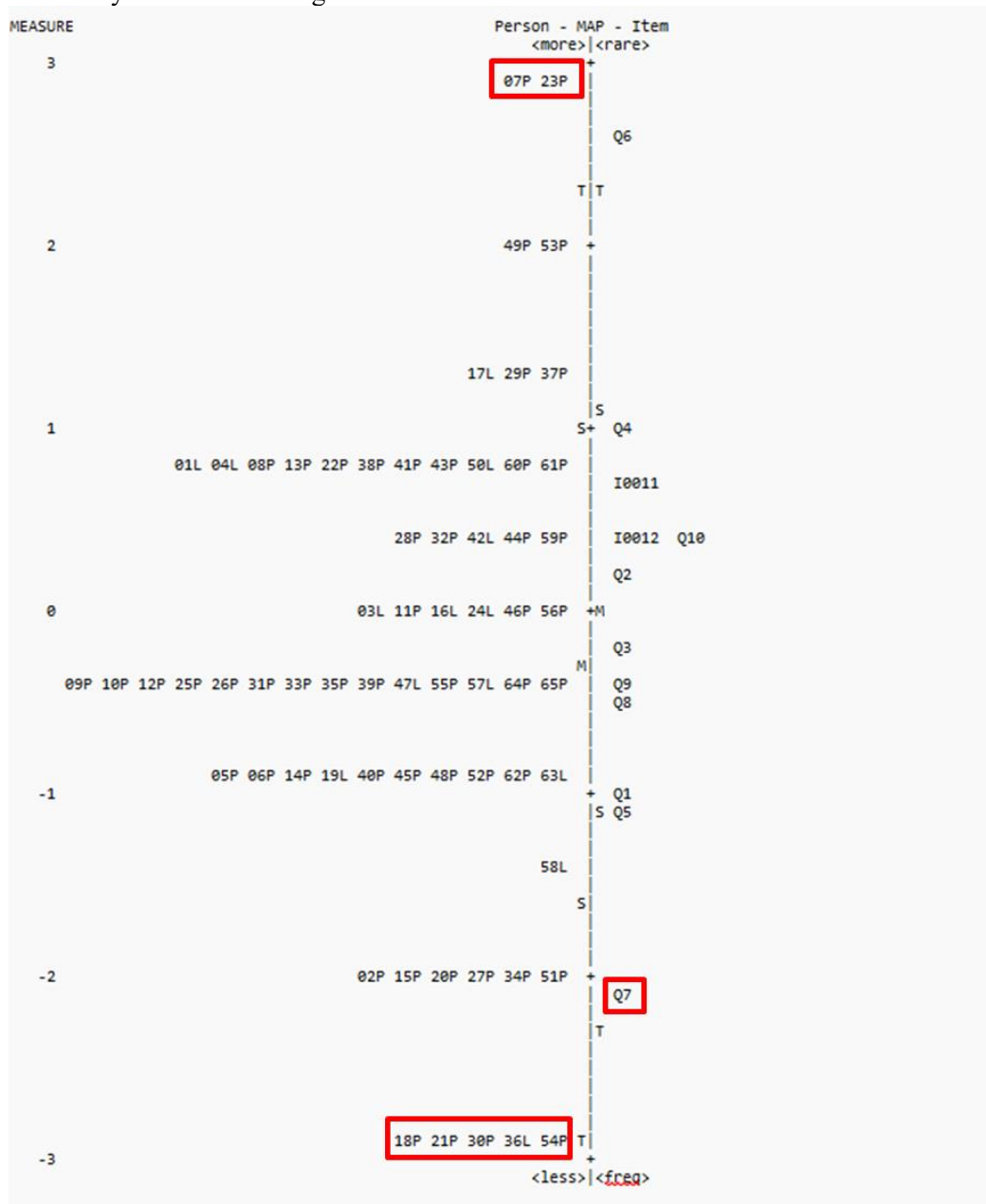


Figure 8. Wright map

The Wright Map showed that students' scientific reasoning abilities were broadly distributed across the logit scale, ranging from approximately (-3 to +3) logits, indicating substantial heterogeneity in reasoning ability within the sample. Students with higher ability levels were generally able to solve medium- to high-difficulty items, whereas students with lower abilities were only able to answer relatively easy items correctly. Items Q1–Q4, representing proportional reasoning, were generally categorized as low- to medium-difficulty items and could be answered correctly by most students, suggesting that proportional reasoning functioned as one of the more basic and accessible dimensions in this sample a pattern broadly consistent with Sapia et al. (2022) Rasch-based ordering of LCTSR dimensions among Italian secondary students, where proportional reasoning items similarly clustered at the lower end of the difficulty hierarchy,

although their sample showed a comparatively narrower spread of difficulty (approximately  $-1.5$  to  $+1.5$  logits) than observed.

In contrast, item Q6, representing the control-of-variables dimension, occupied the highest position on the Wright Map and was identified as the most difficult and most misfitting item. According to Lawson's framework, scientific reasoning is theoretically expected to progress from proportional reasoning toward more advanced forms of formal reasoning, including control of variables, combinatorial reasoning, and probabilistic reasoning. The empirical ordering observed here partially departs from this expectation: several combinatorial reasoning items (e.g., Q7) were empirically easier than the control-of-variables item, indicating that students in this sample found structured, enumerable tasks more tractable than tasks requiring the isolation of causal variables. This pattern echoes findings reported by Van Vo and Csapó (2021) and Zhou et al. (2016), who likewise identified control-of-variables reasoning as the most cognitively demanding dimension across diverse student populations, reinforcing (Kuhn, 2007), argument that coordinating multiple variables under uncertainty places a distinct burden on working memory and hypothesis-testing skills that combinatorial enumeration does not require to the same degree. At the same time, the magnitude of this difficulty gap in the present study with Q6 positioned more than 1.6 logits above the next most difficult item appears larger than the gaps reported in some of these prior studies, which may reflect either a genuine feature of this sample's instructional exposure to experimental design, or an artifact of Q6's documented misfit, a possibility that should temper how strongly this gap is interpreted.

Probabilistic reasoning items occupied moderate-to-high difficulty positions on the map, suggesting that probabilistic reasoning and evidence-based prediction were not yet fully mastered by many students in this sample. This is broadly consistent with developmental accounts suggesting that probabilistic reasoning, like control of variables, depends on coordinating multiple sources of uncertainty rather than following a single deterministic rule (Gopnik, 2012). However, unlike some prior work that has treated the LCTSR's six original dimensions as following a single, mostly linear hierarchy (e.g., Bao et al., 2022), the present findings together with those of Medová et al. (2020), who also reported non-monotonic difficulty orderings among formal-reasoning items suggest that the relative ordering of these dimensions may be more sample- and context-dependent than the theoretical hierarchy implies, rather than being a fixed property of the construct itself. This distinction matters for interpretation: it suggests that the apparent "non-linearity" found here should be read as a feature of this sample's reasoning profile, not necessarily as a challenge to Lawson's framework as a whole.

## CONCLUSION

This study evaluated the psychometric quality of a scientific reasoning instrument adapted from the LCTSR using the Rasch measurement model. The results indicate functional but imperfect measurement quality: person reliability (0.63) was moderate-to-weak and item reliability (0.92) was strong, while one item (Q6) showed clear misfit and two items (Q4, I0011) showed weaker-than-ideal discrimination. Person ability ranged from  $-2.85$  to  $+2.93$  logits, and the Wright Map showed a reasonably good, though imperfect, alignment between item difficulty and student ability. Proportional reasoning items were generally easier, while probabilistic reasoning and control-of-variables items were more difficult with control-of-variables emerging as the most cognitively demanding dimension, suggesting that students in this sample struggled most with causal reasoning and systematic variable manipulation.

Notably, several combinatorial reasoning items were empirically easier than control-of-variables items, suggesting the development of scientific reasoning does not always follow a linear theoretical progression. As this study used a cross-sectional design, this finding should be read as a snapshot of this sample's reasoning pattern rather than developmental evidence. Accordingly, its instructional relevance is best framed as a potential implication greater emphasis on inquiry-based, causal reasoning activities may be beneficial, rather than a firm recommendation, pending confirmation through instructional or longitudinal research.

The findings also have important implications for physics education. Scientific reasoning constitutes a foundational component of physics learning because understanding physical phenomena requires students to formulate hypotheses, identify relationships among variables, evaluate evidence, and draw logical conclusions. The present results suggest that students demonstrated relatively stronger performance in proportional reasoning than in control-of-variables and probabilistic reasoning, indicating that higher-order reasoning processes remain a challenge. Consequently, physics instruction may benefit from greater emphasis on inquiry-based activities, experimental investigations, and problem-solving tasks that explicitly engage students in identifying, controlling, and evaluating variables within physical systems. Furthermore, the instrument provides a potentially useful diagnostic tool for identifying students' reasoning strengths and weaknesses, enabling teachers to design targeted instructional interventions. By integrating scientific reasoning assessment into classroom practice, physics educators may be better positioned to support the development of deeper conceptual understanding and evidence-based thinking, which are essential competencies for learning physics and for scientific literacy more broadly. In this regard, scientific reasoning assessment should be viewed not merely as an evaluation tool but as a means of informing instructional decisions aimed at fostering students' conceptual and epistemic development in physics learning.

Several limitations should be acknowledged. The instrument was translated and reviewed by two expert judges, but without separate readability testing or piloting, which may limit how fully its content and language validity can be ascertained. The sample was also purposively drawn from a single school, constraining generalizability, while the unbalanced gender distribution limited the strength of the DIF analysis under which item Q7 still showed a significant DIF contrast worth further investigation. Item Q6, despite showing clear misfit, was retained in the analysis rather than removed, a decision noted here as a limitation to be revisited in future work with larger samples. Future studies would benefit from larger and more balanced samples across multiple schools, readability and pilot testing during instrument adaptation, and consideration of testlet modeling given the local item dependence observed between Q4 and Q6.

### **CONFLICTS OF INTEREST**

The authors declare that there is no conflict of interest.

### **AUTHOR CONTRIBUTIONS STATEMENT**

Conceptualization, R.; methodology, R.; investigation, R.; formal analysis, R.; writing, original draft preparation, R.; writing, review and editing, D.S. and W.L.; supervision, R.E., A.S., and I.K. All authors have read and agreed to the published version of the manuscript.

### **AI DISCLOSURE STATEMENT**

The author used ChatGPT during the preparation of this work for language editing, translation, improving academic writing, and enhancing the clarity of the manuscript. After using the tool, the author thoroughly reviewed and edited the content as needed and takes full responsibility for the content of the publication.

### **ACKNOWLEDGMENTS**

The author would like to express sincere gratitude to the school, teachers, and students who supported and participated in the research. Their assistance and cooperation during the data collection process contributed significantly to the completion of this study.

### **FUNDING**

This research received no specific external research funding. The author's graduate study was supported by the Beasiswa Indonesia Bangkit (BIB), a scholarship collaboration between the Ministry of Religious Affairs of the Republic of Indonesia and the Indonesia Endowment Fund for Education (LPDP) under the Ministry of Finance of the Republic of Indonesia.

## REFERENCES

- Abate, T., Michael, K., & Angell, C. (2020). Assessment of scientific reasoning: Development and validation of scientific reasoning assessment tool. *Eurasia Journal of Mathematics, Science and Technology Education*, 16(12), em1927.
- Aminudin, A. H., Adimayuda, R., Kaniawati, I., Suhendi, E., Samsudin, A., & Coştu, B. (2019). Rasch analysis of Multitier Open-ended Light-Wave Instrument (MOLWI): Developing and assessing second-years sundanese-scholars alternative conceptions. *Journal for the Education of Gifted Young Scientists*, 7(3), 557–579. <https://doi.org/10.17478/jegys.574524>
- Amiruddin, M. Z. Bin, Samsudin, A., Suhandi, A., Kaniawati, I., COŞTU, B., Aminuddin, A. H., & Kuniawan, F. (2023). Validity and Reliability of the Global Warming Instrument: A Pilot Study Using Rasch Model Analysis. *Jurnal Pendidikan MIPA*, 24(4), 912–922.
- Bachtiar, R. W., Meulenbroeks, R. F. G., & van Joolingen, W. R. (2022). Mechanistic reasoning in science education: A literature review. *Eurasia Journal of Mathematics, Science and Technology Education*, 18(11), em2178.
- Bao, L., Koenig, K., Xiao, Y., Fritchman, J., Zhou, S., & Chen, C. (2022). Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research*, 18(1), 10115.
- Boone, W. J., Staver, J. R., Yale, M. S., Boone, W. J., Staver, J. R., & Yale, M. S. (2014a). Some Wright map nuances: How to set the probability of success at 65%(or whichever percentage you wish to choose). *Rasch Analysis in the Human Sciences*, 255–271.
- Boone, W. J., Staver, J. R., Yale, M. S., Boone, W. J., Staver, J. R., & Yale, M. S. (2014b). Wright maps: First steps. *Rasch Analysis in the Human Sciences*, 111–136.
- Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The evidence-based reasoning framework: Assessing scientific reasoning. *Educational Assessment*, 15(3–4), 123–141.
- Chen, C.-T., & She, H.-C. (2015). The effectiveness of scientific inquiry with/without integration of scientific reasoning. *International Journal of Science and Mathematics Education*, 13(1), 1–20.
- Choowong, K., & Worapun, W. (2021). The Development of Scientific Reasoning Ability on Concept of Light and Image of Grade 9 Students by Using Inquiry-Based Learning 5E with Prediction Observation and Explanation Strategy. *Journal of Education and Learning*, 10(5), 152–159.
- Depaoli, S., Tiemensma, J., & Felt, J. M. (2018). Assessment of health surveys: fitting a multidimensional graded response model. *Psychology, Health & Medicine*, 23(sup1), 1299–1317.
- Díaz, C., Dorner, B., Hussmann, H., & Strijbos, J.-W. (2023). Conceptual review on scientific reasoning and scientific thinking. *Current Psychology*, 42(6), 4313–4325.
- Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report)*.
- Giri, V., & Paily, M. U. (2020). Effect of scientific argumentation on the development of critical thinking. *Science & Education*, 29(3).
- Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337(6102), 1623–1627.
- Han, J. (2013). *Scientific reasoning: Research, development, and assessment*. The Ohio State University.
- Kambeyo, L., & Csapó, B. (2018). Scientific reasoning skills: A theoretical backgrounds to science education. *Reform Forum: Journal for Educational Research in Namibia*, 26(1), 27–36.
- Khan, K. A., & Rana, R. A. (2021). Effect of Higher Order Thinking Instructional Model on Scientific Reasoning of Grade VIII Students. *Bulletin of Education and Research*, 43(1), 77–92.
- Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Mit Press.
- Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only

- challenge. *Science Education*, 91(5), 710–726.
- Linacre, J. M. (1999). Understanding Rasch measurement: estimation methods for Rasch measures. *Journal of Outcome Measurement*, 3, 382–405.
- Linacre, J. M. (2011). *A User's Guide to Winsteps: Rasch-Model Computer Programs*, winsteps.com.
- Medová, J., Bulková, K. O., & Čeretková, S. (2020). Relations between generalization, reasoning and combinatorial thinking in solving mathematical open-ended problems within mathematical contest. *Mathematics*, 8(12), 2257.
- Novia, N., Syamsu, S., & Riandi, R. (2018). Student's achievement in Lawson's classroom scientific reasoning (LCTSR): the effect of gender and age on scientific reasoning ability. *International Conference on Mathematics and Science Education of Universitas Pendidikan Indonesia*, 3, 542–547.
- Nurhayati, N., Suhandi, A., Muslim, M., Kaniawati, I., Wahyudi, W., Misbah, M., & Darman, D. R. (2025). Assessing the Knowledge and Skills of Prospective Physics Teachers in Designing 4C Skills-Oriented Learning: Rasch Analysis. *Qubahan Academic Journal*, 5(1), 718–741.
- Omarchevska, Y., Lachner, A., Richter, J., & Scheiter, K. (2022). It takes two to tango: How scientific reasoning and self-regulation processes impact argumentation quality. *Journal of the Learning Sciences*, 31(2), 237–277.
- Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, 15(2), 20111.
- Samsudin, A., Wulandari, N., Suhandi, A., Yusup, M., Supriyatman, S., Aminudin, A. H., Sumarni, W., & Kapıcı, H. Ö. (2025). Augmented Reality-based Rebuttal Texts (ARaRaT) on momentum-impulse: Rasch analysis on students' conceptual change. *Qubahan Academic Journal*, 5(1), 368–387.
- Sapia, P., Napoli, F., & Bozzo, G. (2022). The Lawson's test for scientific reasoning as a predictor for university formative success: a prospective study. *Education Sciences*, 12(11), 814.
- Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences*, 1(2), 201–238.
- Sumintono, B. (2018). Rasch model measurements as tools in assesment for learning. *1st International Conference on Education Innovation (ICEI 2017)*, 38–42.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assesment pendidikan*. Trim komunikata.
- Van Vo, D., & Csapó, B. (2021). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: Evidence of validity and latent predictors of item difficulty. *International Journal of Science Education*, 43(13), 2185–2205.
- Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research*, 14(2), 20104.
- Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., Xiao, H., Fu, Z., & Bao, L. (2016). Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity*, 19, 175–187.