Developing an ESD-Aligned Scientific Literacy Instrument for Pre-Service Physics Teachers: A Pilot Rasch Measurement Study

Qisthi Fariyani[1,2], Ani Rusilowati[1*], Wiyanto[1]

[1] *Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia.*
[2] *Faculty of Science and Technology, Universitas Islam Negeri Walisongo Semarang, Indonesia.*

Email: * rusilowati@mail.unnes.ac.id.

Abstract

Scientific literacy is important for helping students respond to global sustainability challenges. However, existing scientific literacy instruments have not fully integrated Education for Sustainable Development (ESD) and have limited psychometric evidence. This study aimed to develop an ESD-aligned scientific literacy instrument for pre-service physics teachers and examine its preliminary measurement quality using the Rasch model. The study used a research and development design based on the 4D model. The instrument consisted of 45 contextual items related to sustainability issues, including climate change, energy, and environmental problems. Six experts evaluated the content validity of the items using Aiken's V. The instrument was then administered to 30 pre-service physics teachers. The data were analyzed using the Rasch model with Winsteps. The results showed high content validity, with Aiken's V values ranging from 0.89 to 1.00. The Rasch analysis showed a person reliability of 0.80 and a Cronbach's alpha of 0.84. However, item reliability was 0.45 and item separation was 0.90, indicating that the item calibration was still preliminary. The item-fit results showed Outfit MNSQ values of 0.58–1.37 and Outfit ZSTD values of −1.55 to 1.84. Point Measure Correlation values ranged from 0.17 to 0.48, indicating that some items had weaker relationships with the overall measure. The Wright Map showed imperfect alignment between person ability and item difficulty. PCA of residuals showed 16.8% raw variance explained by measures and a first-contrast eigenvalue of 4.40, which did not provide strong evidence of unidimensionality. Overall, the findings provide preliminary evidence that the instrument has potential for assessing ESD-aligned scientific literacy. Further studies with larger and more diverse samples are needed to confirm its measurement quality.

Keywords: Education for Sustainable Development; Rasch Measurement; Scientific Literacy; Sustainability Assessment.

INTRODUCTION

The rapid advancement of the 21st century has introduced increasingly complex global challenges, including climate change, energy crises, environmental degradation, and accelerated social and technological transformations. Such conditions require educational systems to move beyond a sole emphasis on conceptual mastery toward fostering learners' abilities to meaningfully apply knowledge in understanding and addressing real-world problems. Within this context, literacy has emerged as a fundamental competency, as it represents an individual's ability to understand, evaluate, interpret, and critically use information across various life situations. Previous studies have demonstrated that literacy competencies significantly contribute to academic achievement, evidence-based decision-making, critical thinking, and individuals' readiness to participate effectively in knowledge- and technology-driven societies (Allari et al., 2022; Carroll et al., 2025; Fan & Li, 2025; Wong et al., 2023). Consequently, the need for literacy assessment instruments capable of producing accurate, objective, and meaningful measurements has become increasingly important in contemporary educational research.

Scientific literacy occupies a strategic position because it is closely associated with the ability to understand scientific phenomena, interpret data, evaluate evidence-based claims, and make rational decisions regarding scientific issues encountered in everyday life (Istyadji, 2023). Scientific literacy extends beyond the acquisition of scientific concepts and factual knowledge; it also encompasses the ability to apply scientific understanding to solve contextual problems, critically evaluate information, and comprehend the interrelationships among science, technology, society, and the environment (Chang et al., 2024; Zhang et al., 2023). Its importance has become increasingly prominent in the digital era, where learners are exposed to vast amounts of scientific information of varying credibility. The ability to critically evaluate information therefore becomes essential to prevent misconceptions and the acceptance of pseudoscientific claims (O'Brien et al., 2021; Tseng et al., 2021). Previous studies further indicate that scientific literacy is strongly associated with critical thinking, scientific argumentation, digital information evaluation, and decision-making regarding socio-scientific issues (Dong & Kula, 2023; Kruit et al., 2025; Qiao et al., 2024). Nevertheless, scientific literacy levels among learners across different educational stages remain suboptimal and vary considerably across regions and student populations (Baltikian et al., 2024; Coppi et al., 2024; Subali et al., 2023).

The need to strengthen scientific literacy has become increasingly significant with the growing demand for the implementation of Education for Sustainable Development (ESD). The ESD framework emphasizes the development of competencies that enable individuals to think systemically, understand multidimensional relationships among complex issues, evaluate scientific evidence, and make responsible decisions concerning sustainability. Such competencies require strong scientific literacy because sustainability-related issues, including climate change, energy, health, and environmental challenges, are inherently connected to the application of scientific knowledge in authentic contexts. Previous research has demonstrated that environmental and climate-related literacy can enhance learners' awareness, responsibility, and ability to respond effectively to global sustainability challenges (Balan et al., 2025; Escoz-Roldán et al., 2020; Gellermann et al., 2025). Other studies have also shown that integrating sustainability contexts into science learning strengthens the relationship between scientific knowledge and social decision-making (Kaur & Dhiman, 2025; Okada, 2024). These findings suggest that scientific literacy may serve as an epistemic foundation for developing ESD competencies.

The development of scientific literacy assessment instruments is therefore essential to support educational evaluation aligned with twenty-first-century demands. Numerous studies have developed literacy assessment instruments across various domains, including digital literacy, information literacy, health literacy, and scientific literacy (Alom & Ramalingappa, 2025; Doan et al., 2025; Liu et al., 2024; Sari et al., 2025). Scientific literacy instruments have likewise been developed across different educational levels and scientific disciplines (Chang et al., 2024; Coppi et al., 2023; Zhang et al., 2023). However, most existing instruments primarily focus on traditional cognitive dimensions and have not explicitly integrated sustainability contexts into their assessment design (Coppi et al., 2023; Liu et al., 2024; Okada, 2024). Moreover, current instruments continue to face challenges related to construct validity, item quality, and score interpretation (Coppi et al., 2024; Zhang et al., 2023). Although several instruments have incorporated contextual and sustainability-related elements, relatively limited attention has been given to developing an assessment that explicitly aligns scientific literacy with ESD-related competencies for pre-service physics teachers and evaluates item functioning through Rasch measurement. These limitations indicate that scientific literacy assessment development requires stronger measurement approaches capable of providing detailed information regarding instrument quality.

Traditional measurement approaches based on Classical Test Theory (CTT) present several limitations in educational instrument development. Item parameters in CTT are highly dependent on sample characteristics, causing interpretations of instrument quality to vary across respondent groups. Furthermore, CTT primarily emphasizes total scores and overall reliability, thus providing limited information regarding item characteristics, item difficulty levels, and the interaction between respondent ability and item quality (Dudley & Cairney, 2023; Fang & Zhang, 2026). Such limitations become more critical when instruments are intended to measure contextual and potentially multidimensional constructs such as scientific literacy within sustainability contexts.

The use of the Rasch Model has become increasingly relevant in addressing these limitations because Rasch measurement enables the estimation of respondent ability and item difficulty on a common logit scale, thereby producing more objective and stable measurements. The Rasch Model provides comprehensive information regarding instrument quality through analyses of item fit, person reliability, item reliability, separation indices, dimensionality, item hierarchy, and person–item interaction. This approach also facilitates evaluation of item alignment with the intended construct and enables visualization of the relationship between student ability distribution and item difficulty through the Wright Map. These capabilities position Rasch measurement not merely as a statistical technique but as a measurement framework that strengthens substantive and empirical interpretations of instrument quality (Darman et al., 2024; Fauzi et al., 2021; Suwono et al., 2022).

Previous studies have demonstrated that the Rasch Model substantially improves the quality of educational instrument development across various literacy domains. Fauzi et al. (2021) reported that Rasch analysis provides stronger evidence regarding the measurement quality of genetics literacy instruments. Similarly, Darman et al. (2024) demonstrated that the Rasch Measurement Model enables more detailed evaluations of item functioning and reliability in scientific inquiry literacy instruments. Suwono et al. (2022) further emphasized that the Rasch approach provides more comprehensive interpretations of pre-service teachers' scientific literacy perception instruments. The application of the Rasch Model in broader educational instrument development has also shown its effectiveness in objectively evaluating item quality and measurement consistency (Ophoff & Egger, 2021; Rico et al., 2026; Watson et al., 2025). Nevertheless, studies focusing on ESD-aligned scientific literacy instruments evaluated through Rasch measurement remain relatively limited, particularly among pre-service physics teachers.

Pre-service physics teachers hold a strategic role in supporting the transformation of science education toward sustainability-oriented practices. Their ability to understand, evaluate, and communicate sustainability-related scientific issues will substantially influence the quality of future science learning. Therefore, developing instruments capable of measuring scientific literacy within the context of Education for Sustainable Development is essential for evaluating pre-service teachers' readiness to address twenty-first-century educational challenges. Such instruments should not only measure mastery of scientific concepts but also assess the ability to evaluate evidence-based information, interpret scientific data, and make informed decisions regarding sustainability issues.

Despite the growing literature on scientific literacy and sustainability oriented assessment, a measurement gap remains regarding instruments that bring these elements together for pre-service physics teachers and provide item-level evidence of measurement functioning. The present study addresses this gap by developing an ESD-aligned scientific literacy instrument and examining its preliminary measurement properties through Rasch analysis. The study therefore focuses not on establishing definitive validation, but on obtaining initial evidence concerning content validity, person-level reliability, item functioning, item difficulty, dimensionality, and person item targeting.

Therefore, this study aims to develop a scientific literacy instrument aligned with Education for Sustainable Development competencies for pre-service physics teachers and to evaluate its preliminary measurement quality using a Rasch measurement approach. This approach is employed to obtain more comprehensive evidence regarding measurement quality through analyses of reliability, item fit, dimensionality, item hierarchy, and person–item interaction. The study is intended as a pilot calibration stage; therefore, the findings are interpreted as preliminary evidence rather than as definitive evidence of instrument validity. The findings of this study are expected to contribute theoretically to the advancement of ESD-aligned scientific literacy assessment while providing a contextualized instrument that can be further refined and calibrated to support science education evaluation in the sustainability era.

Methods

A systematic methodological framework was employed in this study to ensure the development of a valid and reliable scientific literacy instrument aligned with Education for Sustainable Development (ESD) competencies. The methodological procedures consisted of research design, participant selection, instrument development, data collection procedures, and data analysis techniques.

Research Design

This study employed a research and development (R&D) approach by adopting the Four-D (4D) model consisting of four stages: define, design, develop, and disseminate (Thiagarajan, 1974). The 4D model was selected because it provides a systematic framework for developing educational instruments that are both valid and feasible for implementation.

The define stage involved a needs analysis and literature review related to scientific literacy, Education for Sustainable Development (ESD), and scientific literacy assessment. The design stage included the development of the assessment blueprint, indicator formulation, and item construction. The develop stage consisted of expert validation, instrument revision, limited testing, and empirical evaluation using a Rasch measurement approach. Finally, the disseminate stage involved the preparation of the final instrument and the reporting of research findings.

Participants

The participants of this study were pre-service physics teachers enrolled at a university in Indonesia, selected through a purposive sampling technique. This group was chosen based on the strategic role of pre-service physics teachers in supporting the implementation of sustainability-oriented science education. The study was conducted through 30 pre-service physics teachers. Although the sample size was relatively small, Linacre et al. (1994) suggested that a sample of approximately 30 respondents can provide stable preliminary item calibration estimates within ±1 logit at a confidence level of approximately 95%, making it suitable for pilot testing in instrument development studies. Because the present study was designed as a pilot calibration, the sample was used to obtain preliminary evidence of item and person functioning rather than to establish definitive population-level psychometric parameters. The limited sample size and purposive sampling approach should therefore be considered when interpreting item reliability, item separation, and the generalizability of the findings.

Instrument Development

The instrument was developed to assess scientific literacy within the context of Education for Sustainable Development by integrating competencies related to understanding scientific concepts, interpreting data, evaluating evidence-based information, and making decisions regarding sustainability issues. The construct development was based on scientific literacy frameworks emphasizing the application of scientific knowledge in real-life contexts (Istyadji, 2023; Zhang et al., 2023).

Instrument indicators were developed based on key ESD competencies, including critical thinking, systems thinking, evidence-based reasoning, and responsible decision-making. The instrument items were constructed using contextual stimuli representing sustainability-related issues such as climate change, energy, and environmental challenges (Coppi et al., 2023; Okada, 2024). Prior to empirical testing, content validity was evaluated by experts in physics education and educational assessment to determine indicator appropriateness, content relevance, and language clarity.

The resulting assessment comprised 45 contextual items. The items were designed to represent the intended scientific literacy construct through sustainability-related contexts rather than to measure isolated scientific knowledge. The blueprint linked the scientific literacy indicators with ESD-related competencies and contextual sustainability issues, thereby providing a basis for evaluating the content relevance of each item.

Content validity was evaluated by six expert validators with expertise relevant to physics education and/or educational assessment. The validators evaluated the relevance, clarity, and appropriateness of the items using a structured rating procedure. Aiken's V was calculated for each item to quantify the degree of agreement among validators regarding item content relevance.

Data Collection Procedure

Data were collected through test administration using the developed instrument under controlled conditions. Participants' responses were coded into a dichotomous format according to the requirements of Rasch measurement analysis and were subsequently examined for completeness before being analyzed using Winsteps software. The empirical administration was conducted as a limited pilot test. No attempt was made to generalize the observed ability distribution to all pre-service physics teachers; instead, the responses were used to examine the preliminary functioning of the items and the consistency of person measurement.

Data Analysis

Data analysis was performed using the Rasch Model with the assistance of Winsteps software to comprehensively evaluate the measurement properties of the instrument (Darman et al., 2024; Fauzi et al., 2021). The analyses included summary statistics, item fit, item difficulty, Principal Component Analysis (PCA) of residuals, and Wright Map analysis.

Summary statistics were used to examine person reliability, item reliability, separation indices, and Cronbach's alpha coefficients. Reliability values were interpreted according to the following criteria: low (<0.67), acceptable (0.67–0.80), good (0.81–0.90), and excellent (>0.90). Separation values greater than 2.00 indicated adequate discrimination ability (Bond, 2015; Sumintono & Widhiarso, 2015).

Item fit was examined using Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt Measure Corr). Outfit MNSQ values between 0.50 and 1.50 and Outfit ZSTD values between −2.00 and +2.00 were used as primary indicators of item fit (Sumintono & Widhiarso, 2015). Pt Measure Corr was additionally examined as a diagnostic indicator of the direction and strength of the relationship between item responses and the overall measured construct. Because the observed Pt Measure Corr values were below the initially specified 0.40 threshold for some items, they were not treated as a sole basis for rejecting items; instead, their values were interpreted together with MNSQ, ZSTD, and substantive item considerations. This decision rule was adopted because applying the Pt Measure Corr range of 0.40–0.85 as an absolute criterion would classify several items as problematic despite their acceptable MNSQ and ZSTD values. Consequently, item functioning was judged holistically rather than through a single statistic.

Item difficulty analysis was conducted using item measure values expressed in logit units to determine the hierarchy of items from the easiest to the most difficult. Construct validity was further examined through Principal Component Analysis (PCA) of residuals by considering the raw variance explained by measures and the unexplained variance in the first contrast (Linacre et al., 1994). The commonly used reference values were treated as diagnostic guidelines rather than as absolute pass fail criteria. In particular, values of raw variance explained by measures above 20% and a first-contrast eigenvalue below 3.00 were considered indicators of stronger unidimensionality; deviations from these values were interpreted as evidence requiring further examination rather than automatic evidence of instrument failure.

The final stage of analysis involved interpreting the Wright Map to examine the distribution of respondents' abilities and item difficulty levels on the same logit scale, thereby providing insights into the alignment between respondent abilities and item characteristics (Bond, 2015; Sumintono & Widhiarso, 2015).

Results and Discussion

This section presents and discusses the findings concerning the development and measurement quality of the scientific literacy instrument aligned with Education for Sustainable Development (ESD) competencies. The results are organized sequentially according to the instrument development process, including the developed product characteristics, content validity evaluation using Aiken's V, and empirical measurement analysis through Rasch modeling comprising summary statistics, item fit, dimensionality assessment, item hierarchy, and person–item interaction.

Development of the Scientific Literacy Instrument

This study produced an Education for Sustainable Development (ESD)-based scientific literacy instrument for pre-service physics teachers consisting of 45 contextual items. The instrument was developed using the 4D model through the define, design, develop, and disseminate stages. The instrument was designed to assess competencies related to understanding scientific concepts, interpreting data, evaluating evidence-based information, and making decisions regarding sustainability-related issues. Sustainability contexts were integrated into the instrument through stimuli based on issues such as climate change, energy, environmental challenges, and natural resource utilization.

The development of instruments incorporating sustainability contexts has become increasingly important because scientific literacy in the twenty-first century is no longer conceptualized merely as mastery of scientific concepts but also as the ability to apply scientific knowledge critically and responsibly to address social and environmental challenges (Gellermann et al., 2025; Istyadji, 2023; Okada, 2024). This perspective is consistent with the evolution of contemporary literacy assessment frameworks that emphasize the interconnectedness among scientific reasoning, socio-scientific issues, and sustainability competencies within science education (Chang et al., 2024; Kruit et al., 2025). Moreover, ESD-oriented science learning increasingly requires assessment systems capable of capturing learners' competencies in interpreting scientific information and making evidence-based decisions in authentic contexts (Balan et al., 2025; Kaur & Dhiman, 2025).

The developed instrument was also designed using contextual stimuli to generate more authentic measurements. The use of real-world contexts in literacy assessment has been shown to improve measurement relevance and provide more representative insights into learners' abilities to apply scientific knowledge in practical situations (Qiao et al., 2024; Sari et al., 2025). Previous studies have further suggested that contextualized assessments can facilitate the evaluation of higher-order cognitive processes, including critical thinking, scientific reasoning, and decision-making competencies (Chang et al., 2024; Dong & Kula, 2023). These findings indicate that the developed instrument not only evaluates conceptual learning outcomes but also assesses learners' scientific reasoning abilities within sustainability contexts.

Content Validity of the Instrument

The content validation results indicated that all instrument items achieved Aiken's V values ranging from 0.89 to 1.00 based on evaluations from six expert validators. These values suggest that each item demonstrated a very high level of relevance, clarity, and appropriateness with respect to the intended construct. Thus, the content-level evidence supported the inclusion of all 45 items in the subsequent pilot administration.

The high content validity values suggest that scientific literacy indicators integrated with ESD competencies were successfully translated into contextual assessment items. Content validity represents a fundamental aspect of instrument development because it determines the extent to which assessment items adequately represent the target construct (Dudley & Cairney, 2023; Liu et al., 2024). Instruments with strong content validity also have greater potential to generate more accurate and meaningful interpretations of measurement outcomes (Fauzi et al., 2021). Furthermore, high content validity contributes to reducing construct underrepresentation and enhancing the consistency between theoretical frameworks and empirical measurement (Boone et al., 2014).

These findings are consistent with previous instrument development studies indicating that the integration of authentic contexts and competency-based indicators improves construct representation quality in educational assessment (Coppi et al., 2023; Dudley & Cairney, 2023; Sari et al., 2025). Zhang et al. (2023) similarly emphasized that scientific literacy assessment development requires strong alignment among conceptual indicators, assessment contexts, and measurement objectives to represent literacy competencies more comprehensively. Such alignment is particularly important in sustainability-oriented assessments because the construct involves multidimensional competencies extending beyond traditional cognitive domains.

Rasch Measurement Analysis

The measurement quality of the developed instrument was evaluated using the Rasch Model through Winsteps software. The Rasch approach was employed to examine item functioning, measurement reliability, item difficulty levels, construct alignment, and the relationship between respondent ability and item characteristics in a more interpretative and comprehensive manner.

The application of Rasch measurement has become increasingly prominent in educational instrument development because it allows item and person parameters to be estimated on a common logit scale, thereby providing more stable and objective measurement outcomes (Darman et al., 2024; Fauzi et al., 2021). Unlike traditional measurement approaches, Rasch analysis enables detailed evaluations of individual item performance and provides empirical evidence regarding the internal structure of the instrument (Ophoff & Egger, 2021; Suwono et al., 2022).

Summary Statistics

The summary statistics results (Figure 1) showed a person reliability value of 0.80 and a Cronbach's alpha coefficient of 0.84. These values provide encouraging preliminary evidence that the instrument produced reasonably consistent person measurements in the pilot sample.

```
     SUMMARY OF 30 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      15.0      45.0        -.83     .35      1.00    .10    .98    .02 |
|  SEM       1.3        .0         .15     .01       .02    .15    .03    .18 |
| P.SD       7.2        .0         .81     .04       .10    .83    .18    .94 |
| S.SD       7.3        .0         .83     .04       .10    .84    .18    .96 |
| MAX.      34.0      45.0        1.21     .48      1.22   2.42   1.40   2.48 |
| MIN.       5.0      45.0       -2.20     .31       .82  -1.71    .66  -1.65 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .36 TRUE SD     .73  SEPARATION  1.99  Person RELIABILITY  .80 |
|MODEL RMSE    .36 TRUE SD     .73  SEPARATION  2.04  Person RELIABILITY  .81 |
| S.E. OF Person MEAN = .15                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .84  SEM = 2.91
STANDARDIZED (50 ITEM) RELIABILITY = .82
```

FIGURE 1. Summary Statistic of ESD-Aligned Scientific Literacy Instrument

The person separation value was 1.99, which was very close to the reference value of 2.00 but did not exceed the predefined criterion. Therefore, the result should be interpreted as indicating moderate differentiation among respondents rather than as fully adequate separation. In Rasch measurement, person separation reflects the ability of an instrument to distinguish respondents according to different levels of the measured construct. Higher separation values indicate greater discriminatory power and improved differentiation among respondents with varying abilities (Bond, 2015; Sumintono & Widhiarso, 2015).

The item reliability value of 0.45 and item separation value of 0.90 indicated limited stability of the item difficulty hierarchy in the present sample. These results suggest that the item calibration estimates had not yet reached an optimal level of stability. This condition may be attributed to the limited number of respondents, resulting in insufficient variability in participants' ability levels to generate stable item calibrations. Previous studies have demonstrated that sample size and respondent heterogeneity substantially influence the stability of item estimates in Rasch analysis (Coppi et al., 2024; Watson et al., 2025). Linacre et al. (1994) also emphasized that although smaller samples may be acceptable during pilot testing, larger and more heterogeneous samples generally produce more stable item calibrations and stronger measurement precision. Accordingly, the person-level reliability evidence appears more encouraging than the item-level calibration evidence. The latter should be regarded as preliminary and requires confirmation through a larger and more heterogeneous sample. This distinction is important because satisfactory person reliability does not necessarily imply stable item calibration.

These findings are consistent with previous studies employing Rasch measurement in literacy assessment development. Fauzi et al. (2021) reported that acceptable person reliability values indicate an instrument's ability to distinguish respondents according to varying competency levels. Darman et al. (2024) further argued that adequate reliability values reflect consistency in measurement and support more meaningful interpretations of assessment outcomes. In the present study, however, these positive person-level findings should not be generalized as definitive evidence of overall psychometric adequacy because the item-level reliability and separation indices remained limited.

Item Fit Analysis

The item fit analysis results (Figure 2) showed Outfit MNSQ values ranging from 0.58 to 1.37 and Outfit ZSTD values ranging from −1.55 to 1.84. These values were within the specified MNSQ and ZSTD reference ranges, indicating generally acceptable item functioning according to these two statistics. In addition, all items demonstrated positive Point Measure Correlation values ranging from 0.17 to 0.48. However, the Point Measure Correlation values require a more cautious interpretation. The observed range of 0.17–0.48 indicates that although all correlations were positive, several items were below the initially specified reference value of 0.40. Therefore, it would be inappropriate to state that all items satisfied every predefined fit criterion. Instead, the positive correlations indicate that the items generally contributed in the intended direction to the measured construct, while items with relatively low correlations warrant further review and possible refinement.

Item fit analysis plays an essential role in instrument development because it evaluates whether individual items behave consistently with the underlying measurement model (Boone et al., 2014). The acceptable MNSQ and ZSTD statistics provide evidence that the observed response patterns were broadly compatible with the Rasch model. Nevertheless, item fit should be interpreted together with the Pt Measure Corr values and substantive characteristics of the items rather than as evidence that every item is equally strong.

The positive Point Measure Correlation values further indicate that the items contributed positively to the overall measurement construct and were capable of discriminating respondents according to their ability levels. Previous studies have emphasized that well-functioning items are fundamental for establishing valid and reliable educational assessments because misfitting items may introduce measurement error and distort construct interpretation (Darman et al., 2024; Fauzi et al., 2021; Ophoff & Egger, 2021). Accordingly, the current results provide preliminary evidence of acceptable item functioning, while items with lower Pt Measure Corr values should be considered candidates for refinement in subsequent calibration.

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
4	5	30	.98	.51	.98	.05	1.37	.88	A .25	.30	86.7	83.6	L2-1
16	16	30	-1.00	.39	1.14	1.05	1.29	1.67	B .17	.36	63.3	66.7	L3-6
28	11	30	-.22	.41	1.25	1.52	1.28	1.33	C .07	.36	50.0	68.6	L5-1
21	13	30	-.54	.40	1.17	1.14	1.27	1.54	D .15	.37	56.7	67.1	L4-3
36	13	30	-.54	.40	1.12	.88	1.19	1.13	E .21	.37	70.0	67.1	L6-6
11	13	30	-.54	.40	1.14	1.01	1.13	.77	F .21	.37	56.7	67.1	L3-1
1	14	30	-.69	.39	1.10	.75	1.13	.82	G .25	.37	66.7	66.8	L1-1
2	17	30	-1.15	.39	1.09	.66	1.07	.43	H .26	.36	63.3	66.5	L1-2
7	11	30	-.22	.41	1.06	.40	1.09	.50	I .29	.36	70.0	68.6	L2-4
9	9	30	.12	.43	1.03	.24	1.09	.42	J .30	.35	73.3	71.9	L2-6
23	13	30	-.54	.40	1.09	.66	1.09	.56	K .27	.37	63.3	67.1	L4-5
27	11	30	-.22	.41	1.09	.62	1.04	.28	L .27	.36	56.7	68.6	L4-9
45	9	30	.12	.43	.98	-.06	1.09	.41	M .35	.35	73.3	71.9	L8-3
3	10	30	-.05	.41	1.03	.24	1.07	.36	N .31	.36	76.7	69.9	L1-3
35	12	30	-.38	.40	1.01	.13	1.07	.45	O .34	.37	73.3	67.7	L6-5
38	7	30	.51	.46	1.06	.34	1.03	.19	P .28	.33	73.3	77.7	L7-1
43	7	30	.51	.46	1.06	.34	.93	-.08	Q .30	.33	73.3	77.7	L8-1
10	14	30	-.69	.39	1.04	.30	1.01	.09	R .34	.37	60.0	66.8	L2-7
39	17	30	-1.15	.39	1.04	.31	1.00	.03	S .33	.36	63.3	66.5	L7-2
14	10	30	-.05	.41	1.03	.21	.99	.03	T .34	.36	63.3	69.9	L3-4
25	11	30	-.22	.41	1.02	.19	1.03	.22	U .33	.36	70.0	68.6	L4-7
42	11	30	-.22	.41	.98	-.10	1.02	.17	V .37	.36	83.3	68.6	L7-5
13	14	30	-.69	.39	.99	-.03	.97	-.16	W .38	.37	66.7	66.8	L3-3
20	11	30	-.22	.41	.98	-.06	.99	.03	v .38	.36	76.7	68.6	L4-2
18	14	30	-.69	.39	.98	-.07	.96	-.20	u .39	.37	66.7	66.8	L3-8
22	8	30	.31	.44	.98	-.04	.94	-.10	t .37	.34	76.7	74.7	L4-4
24	8	30	.31	.44	.98	-.05	.96	-.05	s .37	.34	76.7	74.7	L4-6
8	8	30	.31	.44	.97	-.08	.87	-.37	r .40	.34	70.0	74.7	L2-5
17	11	30	-.22	.41	.97	-.16	.95	-.16	q .40	.36	70.0	68.6	L3-7
26	12	30	-.38	.40	.97	-.12	.91	-.42	p .41	.37	60.0	67.7	L4-8
29	7	30	.51	.46	.95	-.15	.89	-.21	o .40	.33	80.0	77.7	L5-2
12	8	30	.31	.44	.94	-.21	.84	-.46	n .43	.34	76.7	74.7	L3-2
34	5	30	.98	.51	.89	-.25	.93	.00	m .40	.30	86.7	83.6	L6-4
37	11	30	-.22	.41	.93	-.42	.90	-.46	l .45	.36	76.7	68.6	L6-7
31	5	30	.98	.51	.92	-.15	.85	-.19	k .39	.30	86.7	83.6	L6-1
6	5	30	.98	.51	.91	-.20	.82	-.27	j .41	.30	86.7	83.6	L2-3
30	7	30	.51	.46	.91	-.31	.91	-.17	i .43	.33	80.0	77.7	L5-3
19	9	30	.12	.43	.90	-.46	.88	-.39	h .46	.35	80.0	71.9	L4-1
44	9	30	.12	.43	.90	-.48	.82	-.67	g .48	.35	80.0	71.9	L8-2
32	8	30	.31	.44	.89	-.49	.82	-.54	f .48	.34	76.7	74.7	L6-2
5	9	30	.12	.43	.88	-.61	.78	-.82	e .51	.35	66.7	71.9	L2-2
33	7	30	.51	.46	.87	-.50	.79	-.55	d .49	.33	80.0	77.7	L6-3
15	10	30	-.05	.41	.85	-.88	.80	-.89	c .54	.36	76.7	69.9	L3-5
40	4	30	1.27	.56	.85	-.30	.58	-.72	b .48	.28	86.7	86.6	L7-3
41	5	30	.98	.51	.84	-.45	.75	-.44	a .48	.30	86.7	83.6	L7-4
MEAN	10.0	30.0	.00	.43	.99	.1	.98	.1			72.4	72.3	
P.SD	3.3	.0	.59	.04	.09	.5	.16	.6			9.3	5.8	

FIGURE 2. Item Fit of ESD-Aligned Scientific Literacy Instrument

Several items, such as L3-6 and L4-3, demonstrated relatively lower Point Measure Correlation values than other items. Rather than interpreting these values as evidence that the items necessarily failed to measure the construct, they may indicate that the relationship between responses to these items and the overall measured ability was weaker (Kruit et al., 2025; Qiao et al., 2024). This pattern may arise from differences in item content, contextual interpretation, or the cognitive demands required by particular sustainability-related tasks (Chang et al., 2024; Okada, 2024). Therefore, these items should be examined qualitatively in subsequent revisions, particularly with respect to wording, contextual demands, and alignment with the intended indicator.

Item Difficulty and Wright Map

The item difficulty analysis revealed that the instrument demonstrated a relatively diverse distribution of difficulty levels. Item L7-3 was identified as the most difficult item with a logit value of 1.27, whereas items L1-2 and L7-2 were identified as the easiest items with logit values of −1.15 (Figure 3). This distribution indicates variation in item difficulty, although the interpretation should be made cautiously because item reliability and separation were relatively low in the pilot sample.

A broad range of item difficulty is desirable in educational assessment because it enables the instrument to distinguish respondents across different ability levels more effectively (Bond, 2015). The presence of both easier and more difficult items provides potentially useful coverage of the ability continuum, although the degree to which the item hierarchy can be generalized requires confirmation using a larger sample. Previous studies have similarly emphasized that instruments containing a balanced distribution of item difficulties provide more comprehensive measurement information and improve the accuracy of ability estimation (Fauzi et al., 2021; Suwono et al., 2022; Watson et al., 2025).

The most difficult item (L7-3) may reflect the complexity of the cognitive processes required in sustainability-oriented scientific literacy tasks. Higher difficulty levels are often associated with items requiring the integration of scientific understanding, critical evaluation of evidence, and decision-making regarding contextual sustainability issues (Dong & Kula, 2023; Kruit et al., 2025). Conversely, easier items generally involve more direct interpretation of scientific information or the application of basic conceptual understanding. Because the present analysis does not provide dimension-specific item estimates, however, the difficulty of L7-3 should not be interpreted as evidence that a particular ESD competency was intrinsically more difficult than others.

```
MEASURE Person - MAP - Item
         <more>|<rare>
    1          +  L7-3
             X |T
               |
               |
               |  L2-1   L2-3   L6-1   L6-4   L7-4
               |
              T|
               |
               |S
               |  L5-2   L5-3   L6-3   L7-1   L8-1
               |
               |
    0          +  L2-5   L3-2   L4-4   L4-6   L6-2
             X |
           XXX |  L2-2   L2-6   L4-1   L8-2   L8-3
               |
           XX S|M L1-3   L3-4   L3-5
               |
             X |
            XX |  L2-4   L3-7   L4-2   L4-7   L4-9   L5-1   L6-7   L7-5
               |
            XX |  L4-8   L6-5
             X |  L3-1   L4-3   L4-5   L6-6
               |S
             X |  L1-1   L2-7   L3-3   L3-8
   -1        X +
              M|
               |
               |  L3-6
             X |
               |T L1-2   L7-2
               |
            XX |
               |
        XXXXXX |
               |
          XXX S|
   -2          +
            XX |
               |
               |
               |
               |
             X |
               |
               |
              T|
               |
               |
               |
   -3          +
         <less>|<freq>
```

FIGURE 3. Mapping of Item Difficulty

The Wright Map results indicated an imperfect alignment between respondent ability and item difficulty. The mean person measure was approximately −0.83 logits, whereas the mean item difficulty was set at 0 logits. Thus, the respondent distribution was generally located below the mean item difficulty, indicating that the test was somewhat difficult for the pilot group and that targeting could be improved, particularly by including additional items at lower difficulty levels.

This pattern differs from the earlier interpretation that the instrument was relatively well aligned with respondent ability. Instead, the Wright Map suggests that the current item set provides stronger measurement information for respondents at moderate or higher ability levels than for respondents at the lower end of the continuum. The finding is important for subsequent item development because improved targeting can increase measurement precision across the intended range of respondents.

These findings suggest that pre-service physics teachers' scientific literacy within sustainability contexts still requires further strengthening, particularly in competencies involving higher-order reasoning and evidence-based decision-making. The results are consistent with previous studies conducted by Balan et al. (2025) and Baltikian et al. (2024), which reported that learners' scientific literacy levels generally remain within moderate categories and are not yet optimal for addressing complex scientific contexts. Similarly, Dong & Kula (2023) emphasized that contemporary scientific literacy increasingly involves competencies related to information evaluation, critical interpretation of evidence, and contextual decision-making in real-world situations. These findings further imply that science education should not only focus on conceptual understanding but should also facilitate opportunities for learners to engage in scientific reasoning and authentic problem-solving activities within sustainability contexts.

PCA of Residuals

Construct alignment was examined through Principal Component Analysis (PCA) of Residuals, which revealed a raw variance explained by measures value of 16.8%, whereas the unexplained variance in the first contrast was 4.40 eigenvalue units. These values do not provide strong evidence of unidimensionality according to the reference criteria specified in the Methods. Specifically, the raw variance explained by measures was below 20%, while the first-contrast eigenvalue exceeded 3.00. Therefore, the dimensional structure of the instrument requires further investigation.

The emergence of unexplained variance in the first contrast indicates the presence of residual variation that may be associated with the complexity of the scientific literacy construct within sustainability contexts. This construct involves not only conceptual understanding of scientific principles but also data interpretation, evidence-based information evaluation, scientific argumentation, and decision-making regarding social and environmental issues. The multidimensional content demands embedded within these domains may contribute to the observed residual structure.

These findings are consistent with studies by Gellermann et al. (2025), Okada (2024), and Qiao et al. (2024), which suggest that scientific literacy and sustainability literacy are complex constructs that cannot easily be reduced to a single cognitive dimension. Kruit et al. (2025) further reported that assessments based on socio-scientific issues tend to involve multiple interconnected dimensions of scientific reasoning and argumentation. Similar findings were reported by Coppi et al. (2024), who observed that scientific literacy instruments frequently demonstrate moderate multidimensional residual structures due to the complexity of contextual content and literacy task characteristics.

The presence of acceptable MNSQ and ZSTD statistics alongside the residual dimensionality evidence suggests that item functioning and dimensional structure should be considered as separate aspects of measurement evaluation. Thus, acceptable item-fit statistics alone do not establish unidimensionality. The current results support the use of the instrument as a preliminary assessment prototype, while subsequent studies should investigate its dimensional structure using larger samples and, where theoretically justified, multidimensional measurement approaches.

Scientific Literacy Ability of Pre-Service Physics Teachers

The Wright Map results (Figure 4) indicated that most pre-service physics teachers were categorized within the lower portion of the measured ability continuum relative to the item distribution. Respondents' ability distribution was concentrated predominantly within negative logit values and around the midpoint of the measurement scale. This distribution indicates that many respondents had lower estimated overall ability than the average item difficulty in this pilot administration. However, the Wright Map provides an estimate of overall person ability and does not by itself establish that respondents were specifically weak in scientific information evaluation, data interpretation, or evidence-based decision-making.

Therefore, the previous interpretation that respondents were particularly weak in specific scientific literacy competencies should be treated cautiously. Dimension-specific conclusions would require separate subscale or multidimensional analyses that were not conducted in the present pilot study.

Previous studies have highlighted that higher-order scientific literacy competencies require learners not only to acquire scientific concepts but also to critically analyze information, synthesize evidence, and formulate informed judgments regarding complex socio-scientific issues (Dong & Kula, 2023; Kruit et al., 2025; Qiao et al., 2024). The observed person distribution nevertheless provides useful preliminary information for instrument refinement. The relatively low person mean compared with the item mean suggests that additional easier items may be useful to improve targeting, while maintaining some more difficult items to cover respondents with higher ability. Consequently, strengthening these competencies through sustainability-oriented instructional approaches and contextual learning experiences becomes increasingly important for preparing future teachers to address twenty-first-century educational challenges.

```
MEASURE       Person - MAP - Item
                  <more>|<rare>
    1                   +  L7-3
                     1  |T
                        |
                        |
                        |  L2-1  L2-3  L6-1  L6-4  L7-4
                        |
                       T|
                        |
                        |S
                        |  L5-2  L5-3  L6-3  L7-1  L8-1
                        |
                        |
    0                   +  L2-5  L3-2  L4-4  L4-6  L6-2
                     1  |
               0  0  1  |  L2-2  L2-6  L4-1  L8-2  L8-3
                        |
                  0  1 S|M L1-3  L3-4  L3-5
                        |
                     1  |
                  0  1  |  L2-4  L3-7  L4-2  L4-7  L4-9  L5-1  L6-7  L7-5
                        |
                  1  1  |  L4-8  L6-5
                     1  |  L3-1  L4-3  L4-5  L6-6
                        |S
                     0  |  L1-1  L2-7  L3-3  L3-8
   -1                0  +
                       M|
                        |
                        |  L3-6
                     0  |
                        |T L1-2  L7-2
                        |
                  0  0  |
                        |
      0  0  0  0  1  1  |
                        |
               0  1  1 S|
   -2                   +
                  0  0  |
                        |
                        |
                        |
                        |
                     1  |
                        |
                        |
                       T|
                        |
                        |
                        |
   -3                   +
                  <less>|<freq>
```

FIGURE 4. Mapping of Science Literacy Skills of Pre-service Physics Teachers

The findings of this study provide preliminary evidence that the instrument can differentiate respondents along a range of overall scientific literacy ability, although the person separation value of 1.99 indicates that this differentiation is not yet optimal. Previous research indicates that scientific literacy among students and university learners can remain moderate and that learners may experience difficulties in interpreting and evaluating scientific information (Balan et al., 2025; Baltikian et al., 2024). Such literature provides contextual support for continued attention to scientific literacy development, but the present pilot results should be interpreted primarily as evidence about instrument functioning rather than as a population-level estimate of pre-service teachers' scientific literacy.

This condition may be explained by the nature of scientific literacy within the Education for Sustainable Development (ESD) framework, which requires not only mastery of scientific concepts but also the ability to critically and reflectively connect scientific knowledge with social, environmental, and sustainability-related issues (Gellermann et al., 2025; Kruit et al., 2025; Okada, 2024). Sustainability-oriented scientific literacy therefore extends beyond conceptual understanding and requires learners to apply scientific knowledge in multidimensional and authentic contexts.

The findings of this study are consistent with previous research indicating that scientific literacy among both students and university learners generally remains at moderate levels and has not yet reached optimal performance in addressing complex scientific contexts. Baltikian et al. (2024) reported that learners continue to experience difficulties in interpreting and evaluating scientific information. Similar findings were also reported by Balan et al. (2025), demonstrating that learners tend to perform better on concept-based questions than on tasks requiring scientific argumentation and evidence-based decision-making. These findings suggest that learners frequently possess foundational conceptual understanding but encounter challenges when required to apply such knowledge in complex and contextualized situations.

The suboptimal level of scientific literacy among pre-service physics teachers represents an important concern because this group plays a strategic role in supporting the future implementation of sustainability-oriented science education. Teachers are expected not only to understand scientific concepts but also to facilitate students' abilities to evaluate information, engage in critical thinking, and make informed decisions regarding socio-scientific issues (Chung et al., 2024; Istyadji, 2023; Kruit et al., 2025). Previous studies have also emphasized that future teachers require broader competencies encompassing scientific reasoning, reflective thinking, and sustainability awareness to effectively respond to contemporary educational challenges (Escoz-Roldán et al., 2020; Gellermann et al., 2025). These findings indicate that strengthening sustainability-oriented scientific literacy among pre-service physics teachers remains an important priority in higher education contexts.

The Wright Map results further indicated that several high-difficulty items were positioned above the average respondent ability level. This finding suggests the existence of a gap between the complexity of sustainability-oriented scientific literacy demands and students' actual competencies. However, the presence of difficult items remains essential because such items facilitate the identification of students with higher levels of scientific literacy and provide valuable information regarding competency domains that require further improvement. From a measurement perspective, a balanced distribution of item difficulties is important for ensuring that the instrument captures a broad range of abilities and produces more accurate estimations of respondents' competency levels (Bond, 2015; Sumintono & Widhiarso, 2015).

Limitations

Several limitations should be considered when interpreting the findings. The empirical calibration involved only 30 pre-service physics teachers selected purposively from one university; therefore, the item reliability and separation estimates should be regarded as preliminary and should not be generalized to broader populations. The Wright Map also indicated imperfect person item targeting, suggesting that additional items, particularly at lower difficulty levels, may be needed to improve measurement coverage. In addition, PCA of residuals did not provide strong evidence of unidimensionality, indicating that the dimensional structure of the construct requires further investigation. The present study also focused on preliminary Rasch evidence and did not examine the stability of the instrument across larger and more diverse samples. Future research should therefore employ larger and more diverse samples, examine the dimensional structure more rigorously, and recalibrate the revised item set before making definitive claims about the instrument's psychometric quality.

CONCLUSION

This study developed an ESD-aligned scientific literacy instrument for pre-service physics teachers using the 4D development model and examined its preliminary measurement properties through Rasch analysis. The resulting instrument consisted of 45 contextual items integrating scientific understanding, evidence evaluation, data interpretation, and decision-making related to sustainability issues. Content validation provided strong evidence of item relevance, with Aiken's V values ranging from 0.89 to 1.00. The pilot Rasch analysis showed encouraging person-level reliability and generally acceptable item-fit statistics, although the item reliability and separation indicated that the item hierarchy was not yet stable. The PCA of residuals and Wright Map also indicated that the dimensional structure and person item targeting require further improvement. Overall, the instrument shows potential as an ESD-aligned scientific literacy assessment, but the current evidence should be interpreted as preliminary rather than definitive. Its main contribution lies in integrating sustainability-oriented contexts into scientific literacy assessment while applying an objective Rasch measurement framework for initial item calibration. Further studies with larger and more diverse samples are needed to improve item calibration, strengthen person–item targeting, clarify the dimensional structure, and establish more robust evidence for broader use of the instrument.

REFERENCE

Allari, R. S., Hamdan, K., Albqoor, M. A., & Shaheen, A. (2022). Information literacy: Assessment of undergraduate and graduate nursing students. *Reference Services Review*, *50*(2), 211–221. Scopus. https://doi.org/10.1108/RSR-09-2021-0052

Alom, M. M., & Ramalingappa, V. (2025). Development and Validation of a Digital Literacy Scale and the Evaluation of Post-Graduate Teacher Educators' Digital Literacy. *International Journal of Information and Education Technology*, *15*(10), 2249–2263. Scopus. https://doi.org/10.18178/ijiet.2025.15.10.2421

Balan, H. R., Bulay, M. L., & Raganas, N. S. (2025). Climate science literacy evaluation of senior high school students: A platform for science curriculum enhancement. *International Journal of Evaluation and Research in Education*, *14*(5), 3650–3664. Scopus. https://doi.org/10.11591/ijere.v14i5.33760

Baltikian, M., Kärkkäinen, S., & Kukkonen, J. (2024). Assessment of scientific literacy levels among secondary school students in Lebanon: Exploring gender-based differences. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(3). Scopus. https://doi.org/10.29333/ejmste/14279

Bond, T. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition* (3rd edition). Routledge. https://doi.org/10.4324/9781315814698

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

Carroll, A. J., Borycz, J., Stephens, N. K., & Lowery, A. R. (2025). Scaffolded information literacy and data literacy instruction within undergraduate science and engineering laboratory courses:

A longitudinal assessment. *Journal of Academic Librarianship*, *51*(5). Scopus. https://doi.org/10.1016/j.acalib.2025.103124

Chang, T.-C., Lyu, Y.-M., Wu, H.-C., & Min, K.-W. (2024). Introduction of Taiwanese literacy-oriented science curriculum and development of an aligned scientific literacy assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(1). Scopus. https://doi.org/10.29333/ejmste/13930

Chung, C.-C., Lou, S.-J., & Cheng, Y.-M. (2024). Development and Assessment of Innovative Pedagogical Approaches through a Starlink DBL Course. *International Journal of Engineering Education*, *40*(2), 255–267. Scopus.

Coppi, M., Fialho, I., & Cid, M. (2023). Developing a Scientific Literacy Assessment Instrument for Portuguese 3rd Cycle Students. *Education Sciences*, *13*(9). Scopus. https://doi.org/10.3390/educsci13090941

Coppi, M., Fialho, I., & Cid, M. (2024). Validity evidence based on internal structure of a pilot scientific literacy assessment instrument. *Meta: Avaliacao*, *16*(51), 364–385. Scopus. https://doi.org/10.22347/2175-2753v16i51.4330

Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and Validation of Scientific Inquiry Literacy Instrument (SILI) Using Rasch Measurement Model. *Education Sciences*, *14*(3). Scopus. https://doi.org/10.3390/educsci14030322

Doan, A. T., Nguyen, P. V., Tran, H. T. B., Kongkaew, C., Kristina, S. A., Endarti, D., Sriram, S., & Vo, T. Q. (2025). Vietnamese version of the Health Literacy Questionnaire: Translation, linguistic validation, and assessment of cross-cultural adaptation. *Health Promotion International*, *40*(5). Scopus. https://doi.org/10.1093/heapro/daaf149

Dong, F., & Kula, M. C. (2023). Digital device use and scientific literacy: An examination using Programme for International Student assessment (PISA) 2015 data. *Education Economics*, *31*(3), 288–312. Scopus. https://doi.org/10.1080/09645292.2022.2063797

Dudley, D., & Cairney, J. (2023). How the Lack of Content Validity in the Canadian Assessment of Physical Literacy Is Undermining Quality Physical Education. *Journal of Teaching in Physical Education*, *42*(4), 647–654. Scopus. https://doi.org/10.1123/jtpe.2022-0063

Escoz-Roldán, A., Gutiérrez-Pérez, J., & Meira-Cartea, P. A. (2020). Water and climate change, two key objectives in the agenda 2030: Assessment of climate literacy levels and social representations in academics from three climate contexts. *Water (Switzerland)*, *12*(1). Scopus. https://doi.org/10.3390/w12010092

Fan, W., & Li, Y. (2025). Assessment of postgraduate digital research literacy: Current competencies, desired skills, and enhancement strategies. *Education and Information Technologies*, *30*(15), 21551–21571. Scopus. https://doi.org/10.1007/s10639-025-13622-5

Fang, X., & Zhang, Z. (2026). Examining the Contradictions Between the Theory and Practice of Physical Literacy Assessment: Toward a New Assessment Framework. *Journal of Teaching in Physical Education*, *45*(1), 52–65. Scopus. https://doi.org/10.1123/jtpe.2023-0350

Fauzi, A., Saefi, M., Kristiana, E., Adi, W. C., & Lestariani, N. (2021). Factor and Rasch Analysis on COVID-19 Genetics Literacy Assessment Instrument. *Eurasia Journal of Mathematics, Science and Technology Education*, *17*(11). Scopus. https://doi.org/10.29333/ejmste/11264

Gellermann, D., Michel, H., & Harms, U. (2025). Between Two Worlds: Locating Climate Literacy between Modern Educational Frameworks and Assessment Needs. *Mind, Brain, and Education*, *19*(2), 61–72. Scopus. https://doi.org/10.1111/mbe.70003

Istyadji, M. (2023). Conception of scientific literacy in the development of scientific literacy assessment tools: A systematic theoretical review. *Journal of Turkish Science Education*, *20*(2), 281–308. Scopus. https://doi.org/10.36681/tused.2023.016

Kaur, M., & Dhiman, B. (2025). Performance Evaluation and Scientific Mapping in Sustainability-Focused Green Finance and Literacy. *Anuario Do Instituto de Geociencias*, *48*. Scopus. https://doi.org/10.11137/1982-3908_2025_48_63945

Kruit, P. M., Bredeweg, B., & Nieuwelink, H. (2025). Enhancing students' argumentation skills, content knowledge, and Nature of Science understanding through a web-based educational instrument in the context of socio-scientific issues. *International Journal of Science Education*, *47*(6), 749–768. Scopus. https://doi.org/10.1080/09500693.2024.2348824

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the functional independence measure. *Archives of Physical Medicine and Rehabilitation*, *75*(2), 127–132. https://doi.org/10.1016/0003-9993(94)90384-0

Liu, Y., Wang, J., Zhang, Z., Wang, J., Luo, T., Lin, S., Li, J., & Xu, S. (2024). Development and validation of an instrument for measuring civic scientific literacy. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1). Scopus. https://doi.org/10.1186/s43031-023-00092-3

O'Brien, T. C., Palmer, R., & Albarracin, D. (2021). Misplaced trust: When trust in science fosters belief in pseudoscience and the benefits of critical evaluation. *Journal of Experimental Social Psychology*, *96*. Scopus. https://doi.org/10.1016/j.jesp.2021.104184

Okada, A. (2024). A self-reported instrument to measure and foster students' science connection to life with the CARE-KNOW-DO model and open schooling for sustainability. *Journal of Research in Science Teaching*, *61*(10), 2362–2404. Scopus. https://doi.org/10.1002/tea.21964

Ophoff, J. G., & Egger, C. (2021). Assessment Of German And Austrian Students' Educational Research Literacy: Validation Of A Competency Test Based On Cross-National Comparisons. *Studia Paedagogica*, *26*(4), 27–45. Scopus. https://doi.org/10.5817/SP2021-4-2

Qiao, C., Chen, Y., Guo, Q., & Yu, Y. (2024). Understanding science data literacy: A conceptual framework and assessment tool for college students majoring in STEM. *International Journal of STEM Education*, *11*(1). Scopus. https://doi.org/10.1186/s40594-024-00484-5

Rico, A., Palacios-Agúndez, I., & Gonzalez-Larrañaga, G. (2026). A wikipedia-based Authentic Task About Biodiversity in Pre-service Primary Science Teacher Education: Evaluation, Re-design and Validation of a Multidimensional Rubric for K-12 Content. *Research in Science Education*, *56*(2), 551–578. Scopus. https://doi.org/10.1007/s11165-025-10286-6

Sari, N. K., Mahanal, S., & Lestari, S. R. (2025). Constructing and validating a novel instrument to measure human physiology literacy in biology students. *Journal of Applied Research in Higher Education*, 1–23. Scopus. https://doi.org/10.1108/JARHE-07-2025-0523

Subali, B., Faizah, Z., & Sidiq, M. (2023). Indonesian national assessment support: Can RE-STEM Android app improve students' scientific literacy skills? *International Journal of Evaluation and Research in Education*, *12*(3), 1399–1407. Scopus. https://doi.org/10.11591/ijere.v12i3.24794

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*. Trim Komunikata Publishing House.

Suwono, H., Maulidia, L., Saefi, M., Kusairi, S., & Yuenyong, C. (2022). The Development and Validation of an Instrument of Prospective Science Teachers' Perceptions of Scientific Literacy. *Eurasia Journal of Mathematics, Science and Technology Education*, *18*(1). Scopus. https://doi.org/10.29333/EJMSTE/11505

Thiagarajan, T. (1974). *Instructional Development for Training Teachers of Exceptional Children: A Sourcebook*.

Tseng, A. S., Bonilla, S., & MacPherson, A. (2021). Fighting “bad science” in the information age: The effects of an intervention to stimulate evaluation and critique of false scientific claims. *Journal of Research in Science Teaching*, *58*(8), 1152–1178. Scopus. https://doi.org/10.1002/tea.21696

Watson, E., Gares, S., & Rempel, B. (2025). Is the Course Working? An Account of Our Development of an Instrument to Measure the Science Attitudes and Skills of Undergraduate Students Outside of Science Disciplines. *Teaching and Learning Inquiry*, *13*. Scopus. https://doi.org/10.20343/teachlearninqu.13.29

Wong, G. K.-W., Reichert, F., & Law, N. (2023). Reorienting the assessment of digital literacy in the twenty-first century: A product-lifecycle and experience dependence perspective. *Educational Technology Research and Development*, *71*(6), 2389–2412. Scopus. https://doi.org/10.1007/s11423-023-10278-1

Zhang, L., Liu, X., & Feng, H. (2023). Development And Validation Of An Instrument For Assessing Scientific Literacy From Junior To Senior High School. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1). Scopus. https://doi.org/10.1186/s43031-023-00093-2