JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: https://dx.doi.org/10.15408/jp3i.v15i2.45164
http://journal.uinjkt.ac.id/index.php/jp3i

Validation of a High School Physics Scientific Literacy Test Using Classical Test Theory and Item Response Theory

Nurlina[1], Mutmainna[1,2], Nursakinah Annisa Lutfin[1] , Kaharuddin Arafah[3] , Helmi Abdullah[3]

Program of Study in Physics Education, Universitas Sulawesi Barat, Indonesia[1]
Graduate School of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Indonesia[2]
Department of Physics, Universitas Negeri Makassar, Indonesia[3]

nurlina@unsulbar.ac.id

Abstract

Physics scientific literacy instruments at the senior high school level are commonly validated using only Classical Test Theory (CTT), resulting in limited psychometric evidence. This study aimed to validate the high school physics scientific literacy test (HSP-SLT) using CTT and the Item Response Theory (IRT) framework. The study involved 513 Grade XI and XII students from West Sulawesi Province, Indonesia. The initial instrument consisted of 25 multiple-choice items developed based on the three PISA scientific literacy competencies. CTT analysis was conducted to evaluate item difficulty, discrimination, and internal consistency reliability. Subsequently, IRT analysis using the two-parameter logistic (2-PL) model was performed to estimate item discrimination and difficulty parameters and to examine the test information function to obtain a final instrument with more optimal measurement characteristics. The internal consistency reliability yielded a Cronbach's alpha of 0.50 for the total scale. Item difficulty indices ranged from 0.138 to 0.390, and most items had discrimination indices ≥ 0.20. The initial IRT analysis identified five items with unstable parameter estimates or negative discrimination, which were removed. The final model included 20 items with positive discrimination parameters (0.113-1.027) and more stable difficulty estimates (0.611-6.580). The model fit indices indicated an acceptable fit (RMSEA = 0.019, SRMR = 0.045, CFI = 0.944, TLI = 0.944). The test information function showed that measurement precision was highest at ability levels between θ ≈1 and 2. Integrating CTT and IRT produced a more stable and precise measurement model, particularly for students with moderate to high levels of scientific literacy. The HSP-SLT therefore demonstrates potential as a validated instrument for assessing physics scientific literacy at the senior high school level in Indonesia.

Keywords: classical test theory (CTT), high school physics (HSP), item response theory (IRT), scientific literacy test (SLT), validity.

Abstrak

Instrumen literasi sains fisika di tingkat SMA umumnya divalidasi hanya menggunakan pendekatan classical test theory (CTT), sehingga bukti psikometriknya masih terbatas. Penelitian ini bertujuan untuk memvalidasi tes literasi sains fisika sekolah menengah (HSP-SLT) menggunakan pendekatan CTT dan Item Response Theory (IRT). Penelitian ini melibatkan 513 siswa sekolah menengah kelas XI dan XII di Provinsi Sulawesi Barat. Instrumen awal terdiri atas 25 butir pilihan ganda yang dikembangkan berdasarkan tiga kompetensi literasi sains PISA. Analisis CTT digunakan untuk mengevaluasi tingkat kesukaran, daya pembeda, dan reliabilitas internal. Selanjutnya, analisis IRT menggunakan model logistik dua parameter (2-PL) dilakukan untuk mengestimasi parameter diskriminasi dan kesukaran butir, serta menganalisis fungsi informasi tes guna memperoleh instrumen final dengan karakteristik pengukuran yang lebih optimal. Reliabilitas internal menunjukkan nilai Cronbach's alpha sebesar 0,50. Indeks kesukaran berkisar 0,138–0,390 dan sebagian besar butir memiliki daya pembeda ≥ 0,20. Analisis IRT awal mengidentifikasi lima butir dengan parameter yang tidak stabil atau diskriminasi negatif, yang kemudian dieliminasi. Model akhir terdiri atas 20 butir dengan seluruh parameter diskriminasi positif (0,113-1,027) dan tingkat kesukaran yang lebih stabil (0,611-6,580). Indeks kecocokan model menunjukkan hasil yang memadai (RMSEA = 0,019, SRMR = 0,045, CFI = 0,944, TLI = 0,944). Fungsi informasi tes menunjukkan presisi pengukuran tertinggi pada rentang kemampuan θ sekitar 1-2. Penggunaan CTT dan IRT menghasilkan HSP-SLT dengan karakteristik pengukuran yang lebih stabil dan lebih presisi, terutama pada tingkat kemampuan menengah hingga menengah-atas. Instrumen ini berpotensi digunakan sebagai alat ukur literasi sains fisika yang tervalidasi pada jenjang SMA di Indonesia.

***Kata kunci**: classical test theory (CTT), fisika SMA (HSP), item response theory (IRT), tes literasi sains (SLT), validitas.*

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Introduction

Rapid changes across various aspects and fields require the education system to align with the times (Arohman et al., 2016). To survive in the 21st century, students must possess at least 16 fundamental skills (Coppi et al., 2023; Nursyamsudin & Jaelani, 2021; Shohib et al., 2021), which are key measures of students' learning and reflect the overall quality of a country's education system (Afnan et al., 2023). Among these fundamental skills, basic literacy is a core competency, and scientific literacy is one of the six skills encompassed within this competency (Fives et al., 2014; Nursyamsudin & Jaelani, 2021; OECD, 2015, 2019, 2023). Scientific literacy is a primary objective of science education (Istyadji & Sauqina, 2023; Noor, 2021; Şentürk & Sari, 2018) and involves comprehending and applying fundamental principles. It encompasses mastery of fundamental concepts and the ability to apply this knowledge to solve complex problems, analyse data, and think critically (Sartika et al., 2023, 2024, 2025).

The High School Physics Scientific Literacy Test (HSP-SLT) is a physics education instrument developed by the authors to complement existing scientific literacy instruments (Algiranto, 2022; Indrawati & Sunarti, 2018; Shohib et al., 2021). The essence of developing a scientific literacy instrument that reflects curricular characteristics and local learning practices (Algiranto, 2022; Coppi et al., 2023; Fives et al., 2014; Gormally et al., 2012; Indrawati & Sunarti, 2018; Istyadji & Sauqina, 2023; Rusilowati et al., 2016) represents an effort to complement international-scale assessments such as the Programme for International Student Assessment (PISA) (OECD, 2015, 2019, 2023). This initiative is also grounded in several considerations, including differences in students' backgrounds in Indonesia and abroad, different learning processes, and the scientific phenomena presented in learning (Rusilowati et al., 2016). Several studies have examined the development of scientific literacy assessment in physics in Indonesia. For example, static fluid material (Algiranto, 2022), sound wave material (Indrawati & Sunarti, 2018), and the cycle theme (Rusilowati et al., 2016) were used.

In instrument development, validation processes are critical for enhancing the credibility, reliability, and interpretability of the measurement outcomes. Two primary test development frameworks are classical test theory (CTT) and item response theory (IRT) (Bijlsma et al., 2021; Desjardins & Bulut, 2018; Istiyono, 2020; Raykov et al., 2019). To validate the instrument's quality, we analysed the HSP-SLT using both approaches to address gaps in previous studies (Algiranto, 2022; Indrawati & Sunarti, 2018), which focused solely on CTT. Furthermore, using CTT and IRT approaches is regarded as a comprehensive measurement analysis that can enhance accuracy and provide robust psychometric properties (Ahorsu et al., 2022; A. A. Bichi & Talib, 2018; Shi et al., 2024; Thissen, 2015; Tong et al., 2024; Uyigue & Orheruata, 2019; F. Yang et al., 2020).

Previous studies have extensively explored simultaneous validation using CTT and IRT. For instance, Santoso et al. (2024) used CTT- and IRT-based analyses to evaluate the Light Phenomena Conceptual Assessment (LPCA). Their findings showed alignment between CTT and IRT estimates of LPCA item performance and student ability, supporting the LPCA's utility as an inventory for assessing conceptual understanding of light phenomena across low- to high-ability students (Santoso et al., 2024). CTT and IRT were comparable in estimating the item parameters of the Differential Aptitude Test (DAT). They thus could be used independently or complementarily in developing the DAT (Setiawati et al., 2023).

Similarly, Vilca et al. (2022) evaluated the psychometric properties of the revised Mental Health Inventory-5 (R-MHI-5). They concluded that the scale exhibits robust psychometric evidence from both CTT and IRT perspectives (Vilca et al., 2022). Ly et al. (2021) applied CTT and IRT to investigate Thai participants' vector comprehension using the Test Understanding of Vectors (TUV). They found a strong correlation between the CTT and IRT-derived difficulty indices, affirming the reliability of the TUV (Ly et al., 2021). Eaton et al. (2019) employed both CTT and IRT to evaluate the brief electricity and magnetism assessment (BEMA) and the conceptual survey of electricity and magnetism (CSEM), reporting that the overall difficulty levels of the two tests were comparable. Accordingly, using CTT and IRT together enabled us to provide empirical evidence of instrument validity, positioning the HSP-SLT

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

as an instrument that can be used in physics education, particularly for observing students' scientific literacy.

Conceptual Background

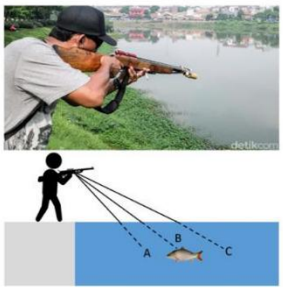
High School Physics Scientific Literacy Test (HSP-SLT)

Scientific literacy is the understanding of scientific concepts and processes needed to make informed personal decisions, participate in public and civic issues, and contribute to economic productivity. It is less about memorising facts and more about applying scientific knowledge and ways of thinking to real-life situations (Shahzadi & Nasreen, 2020). Scientific literacy enables students to understand, identify, explain, and apply scientific findings to address modern society's challenges (Karimah et al., 2021). When addressing real-world problems, scientific literacy fosters scientific thinking skills. It helps individuals respond to societal issues, including developing essential life skills that require sound reasoning within a community context (Oktalia et al., 2021). In physics education, scientific literacy encompasses not only knowledge of physical laws but also the ability to interpret data, evaluate experimental evidence, and apply principles to contextualised problems (Development, 2023; OECD, 2019).

This study instrument consisted of 25 multiple-choice items measuring physics scientific literacy across three core competencies: Explaining Phenomena Scientifically (EPS), Evaluating and Designing Scientific Enquiry (EDSE), and Interpreting Data and Evidence Scientifically (IDES). The EPS dimension assesses students' ability to apply physics concepts to explain natural phenomena and their technological applications. The EDSE dimension evaluates students' capacity to analyse experimental procedures, identify variables, and propose scientifically valid investigations. The IDES dimension measures students' competence in interpreting graphs, tables, experimental results, and evidence-based arguments within physics contexts. By integrating these three competencies, the test aims to capture both conceptual understanding and higher-order reasoning skills that characterise physics scientific literacy at the high-school level.

Figure 1 presents a representative HSP-SLT item that integrates a real-life refraction scenario with a visual ray diagram, requiring students to apply their conceptual understanding to interpret evidence and make an evidence-based decision.

Toni wants to catch fish using an air rifle. The precise location of the fish can be seen in the following pictures and illustrations.



Due to light refraction, the observed fish location is not the actual location. For Toni to successfully aim his air rifle at the fish, the position of the air rifle must be aimed....

A. right at point C
B. right at point B
C. right at point A
D. between points B and C
E. right where there are water bubbles

Sources : detikcom (2020)

Figure 1. Example of HSP-SLT Test Item

http://journal.uinjkt.ac.id/index.php/jp3i

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Classical Test Theory (CTT)

CTT remains a foundational framework for the initial psychometric evaluation of newly developed instruments in educational research. Under CTT, an observed score is interpreted as a function of a true score and measurement error, enabling researchers to estimate reliability and examine item characteristics through indices such as item difficulty, discrimination, and internal consistency (e.g., Cronbach's alpha) (Sharkness & DeAngelo, 2011)(F. Yang et al., 2020). In the context of science education, these indices provide essential preliminary evidence on whether test items adequately differentiate among students with varying proficiency levels, which is critical before more advanced analyses can be interpreted meaningfully (Nurjanah et al., 2024).

For tests measuring complex latent constructs, such as physics scientific literacy, CTT serves as a diagnostic tool to identify problematic items that may not contribute consistently to the overall construct (Petrillo et al., 2015). By evaluating whether items fall within acceptable difficulty ranges and discriminate effectively between high- and low-performing students, CTT screening helps refine test instruments early in development. This is particularly important when the assessment includes contextualised, reasoning-intensive items, where ambiguous or poorly defined content can obscure the construct being measured. Furthermore, recent studies emphasise that classical item analysis remains valuable for ensuring the basic psychometric quality of multiple-choice assessments before latent trait modelling (Nurjanah et al., 2024; Syamsuddin, 2023).

However, despite its utility in early validation, CTT has well-recognised limitations in terms of the stability and generalizability of its estimates. Specifically, item statistics such as difficulty and discrimination depend on the specific sample of test takers, so findings from CTT analyses may not hold across different populations (Eaton et al., 2019; Kong & Lai, 2022; Qian et al., 2025; Shi et al., 2024). Additionally, CTT assumes that measurement precision is uniform across all ability levels (Hartono et al., 2024), a condition that rarely holds for instruments intended to assess broad latent traits, such as scientific literacy. Because of these limitations, CTT is necessary but not sufficient on its own; it must be complemented by frameworks that model latent traits and ability-specific item behaviour. Thus, in validating the HSP-SLT, CTT provides essential baseline psychometric evidence and supports refining the instrument before subjecting it to latent trait analysis via Item Response Theory (IRT).

Item Response Theory

Item Response Theory (IRT) provides a modern, latent trait–oriented framework that extends beyond the limitations of Classical Test Theory by modelling the probability of a correct response as a function of both examinee ability and item characteristics (A. A. Bichi & Talib, 2018; Hays et al., 2000; D. Zhang et al., 2023). IRT allows researchers to estimate item parameters such as difficulty and discrimination independently of the sample on which the test was administered, provided that the model fits the data well (Sharkness & DeAngelo, 2011). This feature enhances parameter stability and supports more generalizable inferences about item quality and student proficiency (Ayanwale et al., 2024). IRT also permits the construction of Item Characteristic Curves (ICCs) and Test Information Functions (TIFs), which together illustrate how measurement precision varies across different levels of ability within the latent trait continuum, which is critical information when validating instruments designed to capture high-order reasoning and multidimensional skills such as scientific literacy.

In educational contexts, especially for assessments such as the HSP-SLT, IRT offers key advantages for understanding how items behave across the full spectrum of student ability. For example, multidimensional IRT models allow for the calibration of items that do not conform to a unidimensional structure, revealing how different latent traits influence item responses (Ayanwale et al., 2024). By leveraging IRT diagnostics, researchers can fine-tune the instrument to ensure it provides reliable, precise measurements for both low- and high-ability students, supporting valid score interpretation for both research and practical educational decision-making.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

The integration of CTT and IRT is complementary rather than redundant (Huang et al., 2020; Qian et al., 2025; Shi et al., 2024). The CTT provides a global overview of reliability and preliminary item screening, ensuring that the test meets fundamental psychometric standards. IRT, in contrast, offers latent trait modelling and ability-specific precision, enabling a deeper analysis of item behaviour. Thus, the dual application of CTT and IRT enhances the instrument's credibility, robustness, and interpretability (D. Zhang et al., 2023; L. Zhang et al., 2023).

Purpose of the Current Study

This study aimed to develop and validate the HSP-SLT using a psychometric approach that integrates CTT and IRT. CTT analysis was employed to evaluate item quality in terms of difficulty, discrimination, and reliability. Subsequently, IRT analysis was conducted to estimate item parameters and examine the test information function to obtain a final instrument with optimal measurement. Based on these objectives, the research questions guiding this study were as follows:

RQ1: What are the item quality and reliability characteristics of the HSP-SLT based on CTT analysis?

RQ2: What are the item characteristics, students' latent ability distribution, and test information function of the HSP-SLT based on IRT analysis?

Methods

Research Design

This study employed a quantitative psychometric validation design to examine the measurement properties of the HSP-SLT developed by the authors. The validation process used two complementary measurement frameworks: CTT and IRT.

Participants

Data were collected from 513 senior high school students (SMA/MA) in West Sulawesi Province, Indonesia. A purposive sampling technique was employed (Denieffe, 2020), based on the following inclusion criteria: (1) students who were enrolled in Grade XI or XII and had studied physics content aligned with the indicators specified in the HSP-SLT blueprint; (2) students who had completed the core physics topics assessed in the instrument; and (3) students representing diverse levels of academic ability.

The sample size of 513 students was determined based on methodological recommendations in psychometric research rather than a conventional statistical power analysis. In the CTT framework, a minimum of 5-10 respondents per item is generally recommended, with a total sample size of at least 300 participants to ensure stable estimates of item validity and reliability. Furthermore, within the IRT framework, particularly for the two-parameter logistic (2-PL) model, approximately 500 respondents are commonly recommended to obtain accurate and stable item parameter estimates (Cappelleri et al., 2014). Therefore, the sample size in this study was considered adequate to produce stable parameter estimates and support a robust psychometric evaluation of the instrument.

Developing the Test Items

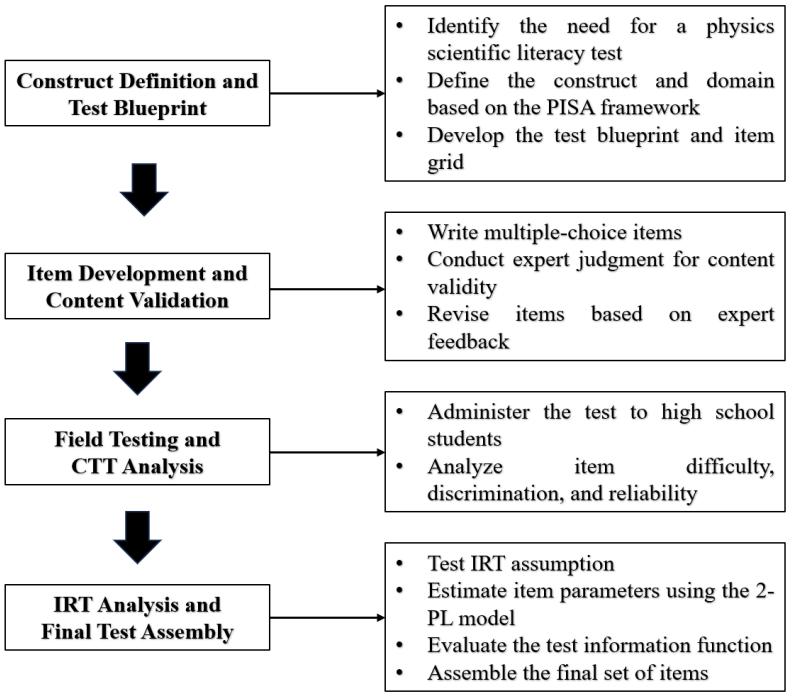
The HSP-SLT was developed systematically by integrating the diagnostic instrument development framework proposed by Adams and Wieman (2011) with general guidelines for item development and content validation grounded in psychometric literature (Boateng et al., 2018; Kalkbrenner, 2021). As illustrated in Figure 2, the instrument development process in this study consisted of four main stages: (1) construct definition and test blueprint development; (2) item development and content validation; (3) field testing and CTT analysis; and (4) IRT analysis and final test assembly.

The indicators of physics scientific literacy in this study were aligned with the PISA framework. Although the items were developed to represent these three competencies, the instrument was designed

http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

to measure physics scientific literacy as a unidimensional construct. Therefore, the resulting score represents the overall scientific literacy ability rather than separate subscale scores for each competency.

tes literasi sains (SLT)



Sources : Personal data.

Figure 2. Instrument Development Procedure for the HSP-SLT

Item development followed a deductive approach informed by a comprehensive literature review, analysis of the senior high school curriculum, and alignment with scientific literacy indicators. In the initial stage, we constructed 42 multiple-choice items to represent all indicators across the three competencies. Three experts evaluated content validity: one physics education lecturer and two educational assessment experts. Experts used a structured rating form to assess indicator alignment, conceptual accuracy of physics content, item construction quality, and linguistic clarity. Each expert rated the relevance of each item using a four-point scale (1 = not relevant to 4 = highly relevant).

The Item Content Validity Index (I-CVI) was calculated to determine the level of interexpert agreement. With three experts, an item was considered content valid if all experts rated it as relevant (I-CVI = 1.00) (Polit & Beck, 2006). The results indicated that 25 items met this criterion, whereas the remaining items were revised or eliminated based on expert feedback. Table 1 shows the distribution of the 25 retained items across the three scientific literacy competencies.

Table 1. Theoretical Framework of the HSP-SLT

Competency of Scientific Literacy	Item Number
Explaining Phenomena Scientifically (EPS)	EPS1, EPS2, EPS3, EPS4, EPS5, EPS6, EPS7, EPS8, EPS9, EPS10, EPS11
Evaluating and Designing Scientific Enquiry (EDSE)	EDSE1, EDSE2, EDSE3, EDSE4, EDSE5, EDSE6, EDSE7, EDSE8
Interpreting Data and Evidence Scientifically (IDES)	IDES1, IDES2, IDES3, IDES4, IDES5, IDES6

Sources : Personal data.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

We administered the content-validated instrument, consisting of 25 items, directly to students. Student responses were coded dichotomously, with 1 indicating a correct response and 0 indicating an incorrect response. This study did not use a separate limited pilot test before the main study. Instead, we conducted a single-stage, large-scale field test using the full sample. This approach enabled iterative item analysis using the CTT and IRT within a single primary dataset. This procedure aligns with contemporary instrument development recommendations, which suggest conducting item reduction directly on the main dataset when the sample size is adequate and psychometric analyses are performed sequentially and systematically (Boateng et al., 2018).

Data Analysis

In this study, data analysis used two primary approaches: CTT and IRT. These approaches were applied complementarily to obtain psychometric evidence from two distinct measurement perspectives. CTT provides sample-dependent descriptive information regarding reliability and item characteristics, whereas IRT offers model-based parameter estimates that are relatively invariant across different samples (C. E. DeMars, 2018; Embretson & Reise, 2013). All analyses were performed using the open-source R statistical software (R version 4.4.2). We used several R packages, including "psych" (Revelle, 2025), "CTT" (Sheng, 2019), "mirt" (Chalmers, 2012), "dplyr" (Wickham et al., 2023), and "ggplot2" (Wickham et al., 2025).

Classical Test Theory (CTT) Analysis

Within the CTT framework, internal consistency reliability was estimated using Cronbach's alpha and the Kuder-Richardson Formula 20 (KR-20), which are equivalent reliability coefficients for dichotomously scored items. Although Cronbach's alpha assumes essential tau-equivalence and unidimensionality (McNeish, 2017), we reported it here as a conventional descriptive index of internal consistency. We interpreted it cautiously alongside the IRT-based evidence. For instruments measuring broad conceptual domains, moderate reliability coefficients may still be considered acceptable because content heterogeneity does not necessarily result in high internal homogeneity (Doran, 1980).

We evaluated item quality under CTT using the difficulty and discrimination indices. The item difficulty index (*P*) was defined as the proportion of students who answered each item correctly and was calculated as follows:

$$P_i = \frac{N_i}{N} \tag{1}$$

where P_i represents the difficulty index for item i, N_i is the number of students who answered item i correctly, and N is the total sample size. Doran (1980) recommends an optimal CTT difficulty range of 0.40 – 0.60 or a wider range of 0.20 – 0.80. In this study, we applied cut-off values of 0.20 and 0.80 to determine the reasonableness of item difficulty from a CTT perspective.

The discrimination index in the CTT reflects an item's ability to distinguish between high- and low-ability students (Adams & Wieman, 2011; Doran, 1980). It was calculated using the upper and lower 27% groups as follows:

$$D = \frac{High\ (27\%) - Low(27\%)}{N} \tag{2}$$

Discrimination values ranged from 0 to 1, with acceptable indices typically falling between 0.20 and 0.60 (Doran, 1980). These thresholds guided item evaluation in the present study.

Item Response Theory (IRT) Analysis

To complement the CTT analysis, this study employed an IRT approach using a two-parameter logistic (2-PL) model. The 2-PL model was selected because it allows simultaneous estimation of item discrimination (a) and item difficulty (b) parameters, making it more flexible than the one-parameter

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

logistic (1-PL) model in capturing variability in item characteristics (Bock & Gibbons, 2021; Embretson & Reise, 2013).

Prior to parameter estimation, the assumption of unidimensionality was examined through Exploratory Factor Analysis (EFA) using eigenvalue analysis, scree plot inspection, and parallel analysis. A unidimensional IRT model assumes that item responses are primarily influenced by a single dominant latent construct (Bock & Gibbons, 2021). Therefore, the substantially larger first factor relative to the subsequent factors served as the basis for applying a unidimensional model. The instrument was considered to satisfy the unidimensionality assumption if one dominant factor exhibited the largest eigenvalue and the scree plot demonstrated a clear drop after the first factor (Eaton et al., 2019).

The assumption of local independence was evaluated using Yen's Q3 residual statistics (Finch & Jeffers, 2016). Residual correlations close to zero indicate that latent constructs adequately explain covariance among items. Additionally, the monotonicity assumption was considered satisfied when each item had a positive discrimination parameter, resulting in an Item Characteristic Curve (ICC) that increased monotonically as latent ability increased (F. M. Yang & Kao, 2014).

This study used the 2-PL model. The 3-PL and 4-PL models were not retained because of convergence issues. The probability function of the 2-PL model is expressed as

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}} \text{ with } i = 1,2,3,\dots n \quad (3)$$

where θ represents latent ability, a_i denotes the discrimination parameter (reflecting the slope of the ICC), b_i represents the difficulty parameter, and D =1.702 is the scaling constant. We considered item difficulty values within the range of -4 to +4 acceptable, and interpreted discrimination values between 0.34 and 1.35 as indicating moderate-to-good discrimination (Baker, 2001). We removed items with negative discrimination or extreme parameter estimates and iteratively re-estimated the model until it converged.

Model fit was evaluated using the M2 statistic along with global fit indices, including the Root Mean Square Error Approximation (RMSEA $\leq$ 0.08), Standardized Root Mean Square Residual (SRMR $\leq$ 0.08), Comparative Fit Index (CFI $\geq$ 0.90), and Tucker-Lewis Index (TLI $\geq$ 0.90) (Eaton et al., 2019). Evaluating the model fit is essential to ensure that the estimated item response functions adequately represent the observed patterns (Bock & Gibbons, 2021). In addition, we examined measurement precision across the ability continuum using the Test Information Function (TIF) (Bock & Gibbons, 2021; Embretson & Reise, 2013).

Results and Discussion

Results

Descriptive Statistics

Table 2 presents descriptive statistics for the HSP-SLT total scores of 513 students (184 male and 329 female) who completed the 25-item instrument. The mean total score was 6.57 (SD = 3.03), with observed scores ranging from 1 to 21, out of a maximum possible score of 25. The distribution of scores was positively skewed (skewness = 1.46), indicating that a large proportion of students obtained low scores. In addition, the distribution was leptokurtic (kurtosis = 3.43), suggesting a more peaked distribution with heavier tails than a normal distribution. These distributional characteristics suggest the test may have been relatively challenging for the sampled population, as reflected in the concentration of scores in the lower ranges.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Table 2. Descriptive Statistics of HSP-SLT

Test Length	Count	Mean	Variance	SD	Skewness	Kurtosis	Min.	Max.	Med.
25	513	6.573	9.155	3.025	1.46	3.43	1	21	6

Sources : Personal data.

Classical Test Theory (CTT)

The internal consistency reliability of the HSP-SLT was estimated using several indices. The instrument demonstrated moderate reliability, with Cronbach's alpha (α = 0.5), KR-20 = 0.50, KR-21 = 0.49, and the split-half reliability = 0.56. This value is deemed acceptable, given the limited number of items and broad range of knowledge assessed (Doran, 1980). Previous studies in science education have emphasised that instruments designed to measure diverse conceptual domains are not necessarily expected to produce high alpha coefficients(Taber, 2018). In this case, the HSP-SLT assesses multiple physics literacy competencies within a single overarching construct. Therefore, high internal consistency (i.e., strong inter-item homogeneity) is not theoretically anticipated. The moderate reliability observed may reflect heterogeneous content coverage rather than measurement instability (Doran, 1980).

Table 3 presents the item difficulty (*P*) and discrimination (*D*) indices for each HSP-SLT item in the study. The difficulty index ranged from 0.138 to 0.390, with a mean of 0.26, indicating that most items were classified as difficult or moderately difficult. No items were categorized as very easy ($P > 0.80$), suggesting that the instrument posed a substantial cognitive challenge to the sampled students. The discrimination index ranged from 0.013–0.419. Most items demonstrated acceptable discrimination ($D \geq 0.20$), indicating an adequate ability to differentiate between high- and low-performing students. However, several items (EPS2, EPS5, EDSE6, IDES2, IDES3, IDES4, and EDSE3) exhibited low discrimination indices and therefore warrant revision or further psychometric evaluation in future studies. In particular, items with very low discrimination (e.g., EDSE6 and IDES4) may require substantial refinement or potential removal, as they contribute minimally to distinguishing students across ability levels.

Table 3. CTT-Based Estimates of Item Difficulty (*P*) and Discrimination (*D*)

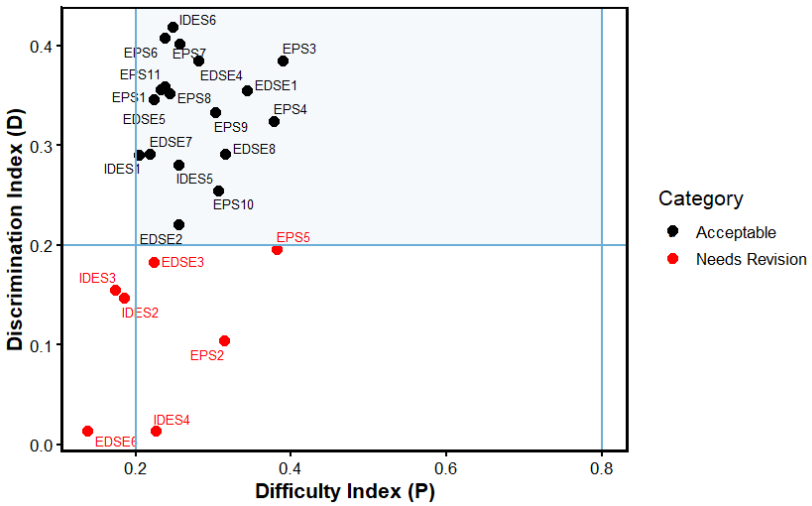
Item	*P*	*D*	Item	*P*	*D*
EPS1	0.23196	0.35567	EDSE3	0.22417	0.18261
EPS2	0.31384	0.10385	EDSE4	0.28070	0.38529
EPS3	0.38986	0.38526	EDSE5	0.22417	0.34652
EPS4	0.37816	0.32423	EDSE6	0.13840	0.01365
EPS5	0.38206	0.19598	EDSE7	0.21832	0.29162
EPS6	0.23781	0.40757	EDSE8	0.31557	0.29155
EPS7	0.25731	0.40176	IDES1	0.20467	0.29061
EPS8	0.24366	0.35205	IDES2	0.18518	0.14701
EPS9	0.30214	0.33307	IDES3	0.17348	0.15496
EPS10	0.30604	0.25470	IDES4	0.22612	0.01313
EPS11	0.23781	0.35910	IDES5	0.25536	0.28086
EDSE1	0.34331	0.35471	IDES6	0.24756	0.41868
EDSE2	0.25536	0.22023			

Sources : Personal data.

http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

As illustrated in Figure 3, most items were clustered within the quadrant representing acceptable ranges of difficulty and discrimination. This pattern indicates that most items functioned adequately under the CTT framework. Only a limited number of items fell outside the recommended discrimination threshold, highlighting the need for refinement.

Overall, the CTT findings indicate that the HSP-SLT demonstrates generally acceptable item functioning, with most items satisfying established difficulty and discrimination criteria. Nevertheless, because the CTT indices are sample-dependent, further evaluation using a model-based approach is necessary. In line with the study's objective of integrating classical and modern psychometric perspectives, the instrument was subsequently analysed using IRT to provide parameter estimates less influenced by sample characteristics and to obtain a more robust evaluation of item performance.



Sources : Personal data.

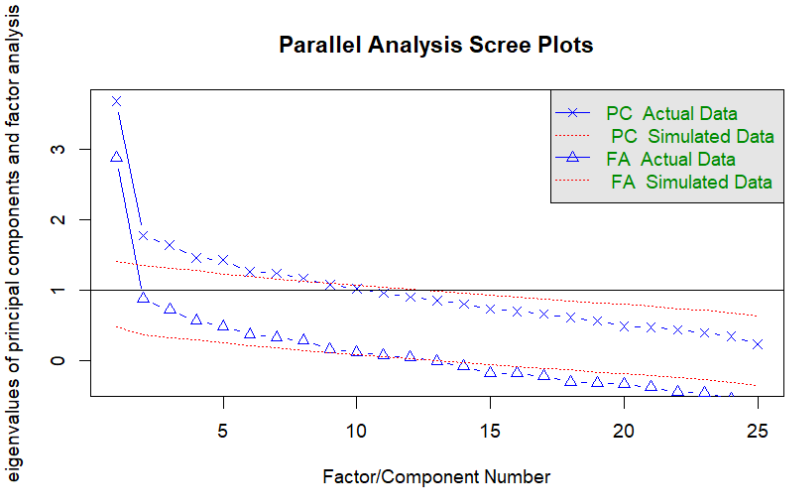
Figure 3. Item Difficulty and Discrimination Indices of the HSP-SLT with CTT

Item Response Theory (IRT)

Before estimating item parameters using the IRT framework, the dimensional structure of the HSP-SLT was examined to ensure the instrument met the assumption of unidimensionality. Dimensionality was evaluated using exploratory factor analysis (EFA). The results indicated that the first factor had an eigenvalue of 3.69, followed by the second (1.77) and third factors (1.64). The ratio between the first and second eigenvalues (2.08) suggests a dominant general factor underlying the item responses. Although the first eigenvalue does not far exceed the subsequent ones, the magnitude difference supports interpreting a primary latent dimension.

Figure 4 presents the scree plot and parallel analysis results. The scree plot shows a clear, steep decline after the first factor, indicating a substantial drop in explained variance. Moreover, the first factor lies well above the simulated eigenvalue line in the parallel analysis, providing further evidence of a strong general scientific literacy factor. Although the second and third factors also exceed the simulated threshold, their relative contribution is much smaller, and the eigenvalue slope becomes more gradual after the third factor.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)



Sources : Personal data.

Figure 4. Scree Plot and Parallel Analysis Results

Taken together, these findings suggest that HSP-SLT possesses a dominant general factor and can be treated as essentially unidimensional. Considering the study's objective of estimating overall scientific literacy ability rather than separate subscale scores, a unidimensional IRT model was selected as the primary model for subsequent analysis.

Attempts to estimate the 3-PL and 4-PL models did not yield stable convergence solutions; consequently, the analysis was restricted to the 1-PL and 2-PL models only. Model comparison based on the Akaike Information Criterion (AIC) showed that the 2-PL model had a lower AIC (AIC = 14,299.19) than the 1-PL model (AIC = 14,310.37). Because lower AIC values indicate a better model fit, the 2-PL model was selected as the primary model for item parameter estimation, as it accommodates variability in the item discrimination parameter across items.

For the assumption of local independence, the results indicated that the mean residual correlation among items was close to zero (Mean = -0.014), indicating no systematic local dependence after controlling for latent ability. Furthermore, the maximum Q3 value was 0.144, which is below the commonly recommended threshold of 0.20 in IRT literature (Christensen et al., 2017). These findings suggest that no item pairs exhibited substantial local dependence. Therefore, the local independence assumption of the 2-PL model can be considered to be satisfied.

Table 4 presents the initial estimates of the discrimination (a) and difficulty (b) parameters based on the unidimensional 2-PL IRT model. The discrimination values ranged from -0.51 to 1.03. Most items showed positive discrimination parameters, ranging from low to high, indicating their ability to distinguish between students with lower and higher levels of scientific literacy.

Table 4. Initial Discrimination (a) and Difficulty (b) Estimates Under the 2-PL IRT Model

Item	a	b	Item	a	b
EPS1	0.71594	1.83586	EDSE3	0.22584	5.55843
EPS2	-0.20749	-3.80901	EDSE4	0.86089	1.23648
EPS3	0.86085	0.57746	EDSE5	0.61557	2.16790
EPS4	0.47198	1.09898	EDSE6	-0.50527	-3.78917
EPS5	0.11521	4.18588	EDSE7	0.49466	2.70769
EPS6	1.02762	1.33699	EDSE8	0.32114	2.46169
EPS7	0.83985	1.42317	IDES1	0.41968	3.35246

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Item	a	b	Item	a	b
EPS8	0.89814	1.44223	IDES2	0.00893	165.85895
EPS9	0.66830	1.35851	IDES3	-0.00291	-535.82205
EPS10	0.43571	1.95338	IDES4	-0.43056	-2.97105
EPS11	0.67696	1.87293	IDES5	0.42184	2.63216
EDSE1	0.66413	1.05689	IDES6	0.99049	1.31131
EDSE2	0.15455	6.96184			

Sources : Personal data.

However, the difficulty parameter (b) showed an extremely wide range, from -538.82 to 165.86. These extreme values do not substantively reflect meaningful variation in item difficulty; rather, they indicate instability in parameter estimation caused by a highly imbalanced response distribution for specific items. Such instability suggests that the initial model calibration was suboptimal and required refinement through the removal of poorly functioning items.

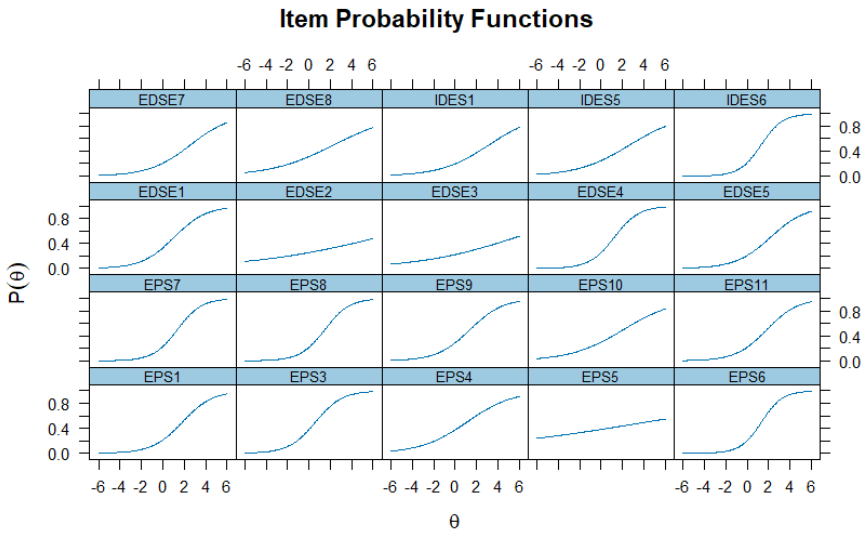
Within the IRT framework, the monotonicity assumption requires that the probability of answering an item correctly increases as latent ability increases, which is reflected in a positive discrimination parameter (Baker, 2001; Embretson & Reise, 2013). The presence of four items with negative discrimination parameters (EPS2, EDSE6, IDES3, and IDES4) violates this assumption, as higher-ability students were less likely to respond correctly. Negative discrimination parameters also suggest potential model misfit or item malfunction (C. DeMars, 2010). Therefore, we removed these four items before re-estimating the model.

After removing the four negatively discriminating items, re-estimation indicated that all remaining items had positive discrimination. Nevertheless, one item (IDES2) demonstrated an extremely low discrimination value (a = 0.040) and an excessively high difficulty estimate (b = 37.487). These characteristics indicate that the item possessed minimal discriminatory power and was rarely answered correctly, resulting in unstable parameter estimates. Consequently, we removed IDES 2 to obtain a more psychometrically robust measurement model.

The 2-PL model was re-estimated using the remaining 20 items. The final estimation results showed that all items had positive discrimination parameters ranging from 0.113 to 1.027, indicating adequate differentiation between students across the ability continuum. The difficulty parameters ranged from 0.611 to 6.570, with most items at moderate to relatively high difficulty levels.

Figure 5 presents the ICC for the final 20-item 2-PL model. All curves displayed a monotonically increasing pattern, confirming that the monotonicity assumption was satisfied and that the model met its theoretical expectations after the refinement process. Overall, the final 20-item model demonstrated substantially more stable and interpretable measurement characteristics than the initial calibration.

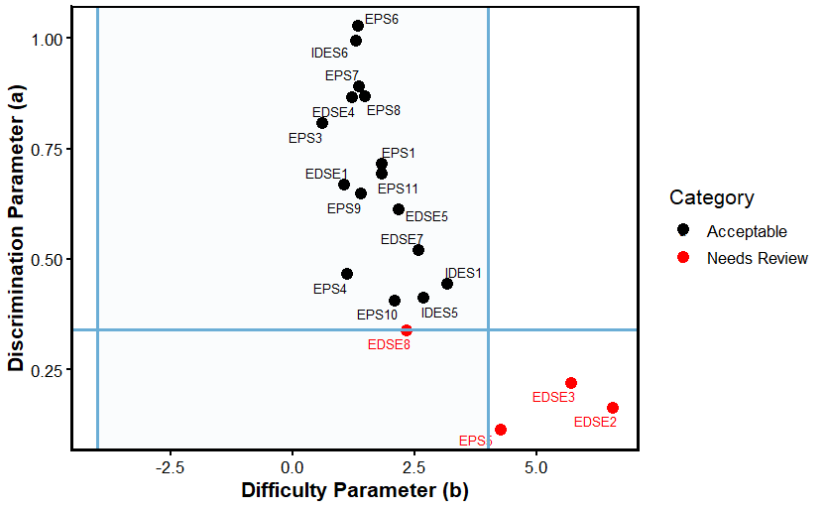
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)



Sources : Personal data.

Figure 5. Item Characteristic Curves (Final 20-Item 2-PL Model)

Figure 6 presents the item characteristic map based on the discrimination and difficulty parameters estimated under the final 20-item 2-PL IRT model of the HSP-SLT. Most items fell within an acceptable range of difficulty and showed discrimination values ≥ 0.34, indicating adequate to good ability to differentiate students across levels of scientific literacy. However, several items demonstrated relatively high difficulty levels (b > 4) or comparatively low discrimination parameters.



Sources : Personal data.

Figure 6. Distribution of Difficulty and Discrimination Parameters (Final 20-Item 2-PL Model)

Table 5 presents the model fit indices of the 2-PL IRT model for each stage of item reduction. The final 20-item model yielded an M2 value of 202.95 with 170 df (p = 0.043). Although the M2 test was statistically significant, it is sensitive to large sample sizes and may lead to rejecting models that otherwise demonstrate an acceptable practical fit. Therefore, we placed greater emphasis on the approximate fit indices, including RMSEA, SRMR, CFI, and TLI. The final model had an RMSEA of 0.019 and an SRMR of 0.045, both indicating excellent fit. Furthermore, the CFI (0.944) and TLI (0.944) exceeded the recommended threshold of 0.90, supporting model adequacy. Taken together, these results indicate that the final 20-item 2-PL model represents the observed response data well and shows substantially improved fit compared to the initial model.

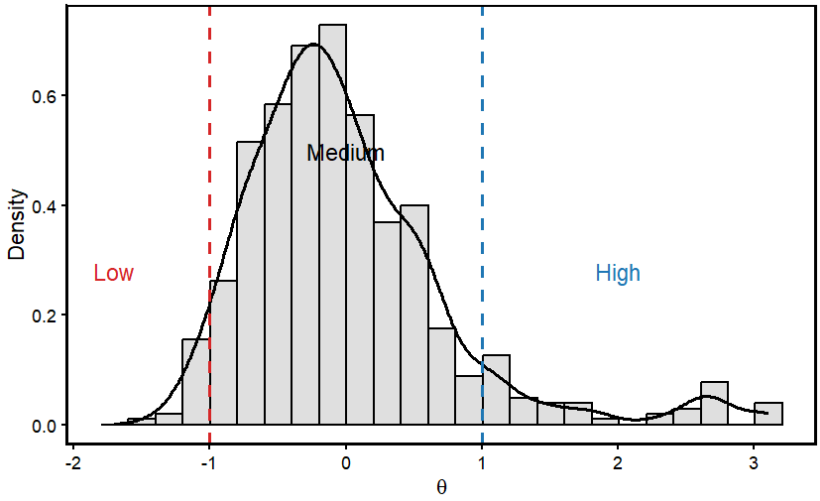
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Table 5. Model Fit Indices for the 2-PL IRT Model Across Item Reduction Stages

Items Remove	Remaining Items	Model Fit Statistics						
		M2	df	p-value	RMSEA	SRMR	CFI	TLI
None	25	334.75	275	0.008	0.021	0.045	0.907	0.898
4 items	21	227.988	189	0.028	0.020	0.045	0.935	0.927
5 items	20	202.95	170	0.043	0.019	0.045	0.938	0.944

Sources : Personal data.

Figure 7 presents the distribution of students' latent ability (θ) estimated using the final 2-PL model. The distribution approximates normality, with a clear peak around $\theta = 0$, indicating that most students cluster within the moderate ability range. A smaller proportion of students is located at the lower and higher ends of the ability continuum. This distribution suggests that the instrument adequately captures variability across a broad range of physics scientific literacy levels.



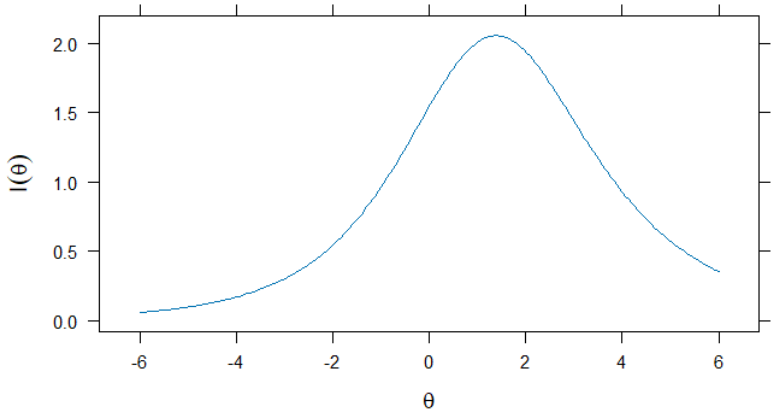
Sources : Personal data.

Figure 7. Distribution of Students' Latent Ability (θ)

Figure 8 illustrates the Test Information Function (TIF), showing that maximum information is achieved at θ between 1 and 2. This indicates that the instrument provides the highest measurement precision for students with moderate to above-average levels of scientific literacy in physics. In the IRT framework, higher test information corresponds to a lower standard error of measurement, reflecting greater reliability of ability estimates within this range. Conversely, the information decreases at the lower and higher ends of the ability level. These results demonstrate that the HSP-SLT is most effective for differentiating students in the middle to upper ranges of scientific literacy.

Overall, the iterative refinement process resulted in a 20-item HSP-SLT with stable item parameters, a satisfactory model fit, and appropriate measurement precision across moderate to high ability levels. These findings indicate that integrating the CTT and IRT approaches strengthened the instrument's psychometric robustness.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)



Sources : Personal data.

Figure 8. Test Information Function

Discussion

The findings indicate that the developed HSP-SLT demonstrates adequate psychometric properties after a systematic and iterative refinement process. The CTT analysis showed that most items fell within the moderate-to-difficult range and exhibited acceptable discrimination indices. Subsequently, the IRT analysis using the 2-PL model resulted in a final 20-item instrument with all discrimination parameters positive, more stable difficulty estimates, and acceptable global model fit indices. The TIF revealed that the highest measurement precision was achieved at ability levels between $\theta \approx 1$ and 2 (El-Hamamsy et al., 2022), indicating that the instrument is most sensitive in differentiating students with moderate to moderately high levels of physics scientific literacy. Substantively, this suggests that the instrument is particularly effective in identifying students transitioning from basic conceptual recall to higher-order reasoning and evidence-based scientific thinking. These findings confirm that the study's objective of developing an instrument supported by validity evidence from both the CTT and IRT perspectives has been successfully achieved.

These results align with previous research showing that combining CTT and IRT approaches provides a more comprehensive psychometric evaluation. While CTT offers an initial overview of overall test quality, IRT enables a more in-depth examination of item characteristics that are less dependent on sample characteristics (Qian et al., 2025; Tong et al., 2024; C. Zhang et al., 2020). In the present study, the consistency between the two approaches supports the value of integrating classical and modern measurement frameworks. The literature recommends combined approaches because they provide stronger, more comprehensive validity evidence than relying on a single framework (Boateng et al., 2018; Cappelleri et al., 2014). This alignment between CTT and IRT findings has also been reported in prior studies examining item characteristics and participant ability estimation (A. Bichi et al., 2019; Santoso et al., 2024; Setiawati et al., 2023), as well as consistency in parameter estimation and reliability across both approaches (Vilca et al., 2022), including in physics education contexts (Eaton et al., 2019).

Several initial IRT findings were unexpected. Negative discrimination parameters in several items indicated that higher-ability students were less likely to answer those items correctly, contradicting the IRT monotonicity assumption. Such patterns typically indicate item construction issues, including ambiguous wording, non-functioning distractors, multidimensional influences, or potential misalignment between the intended cognitive demand and the actual reasoning required. Several IDES items showed unstable results, indicating that tasks requiring students to interpret data and draw evidence-based conclusions need careful design, with clear stimuli and appropriate levels of reasoning demand. IRT literature clearly states that items with negative discrimination should be removed because they do not provide valid information about the latent trait (Embretson & Reise, 2013; F. M. Yang & Kao, 2014). Therefore, removing these items in the present study was methodologically appropriate.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Additionally, the extremely large difficulty estimates in the initial model did not substantively reflect true item difficulty levels. Instead, they reflected parameter-estimation instability due to a highly imbalanced response distribution. This situation commonly occurs when nearly all respondents answer an item correctly or incorrectly, making accurate parameter estimation difficult. After eliminating problematic items and recalibrating the model, the parameter estimates became more reasonable and stable. These findings underscore the importance of iterative recalibration in IRT analysis to obtain a valid and interpretable measurement model.

The moderate internal consistency reliability ($\alpha = 0.5$) should be interpreted cautiously. This value may be influenced by the relatively limited number of items and the broad and heterogeneous nature of the scientific literacy construct being measured. Instruments encompassing multiple conceptual domains may exhibit moderate reliability coefficients without necessarily indicating weak measurement quality (Doran, 1980; Goldmann et al., 2025; Tavakol & Dennick, 2011). Thus, the moderate alpha value in this study more likely reflects content breadth rather than measurement instability. Moreover, the IRT results demonstrated satisfactory measurement precision within specific ability ranges, which can be interpreted as model-based reliability.

A further conceptual consideration concerns applying a unidimensional IRT model to an instrument developed around three PISA scientific literacy competencies. Factor analysis indicated a dominant general factor, supporting the treatment of the instrument as essentially unidimensional for estimating overall scientific literacy ability. This interpretation aligns with prior research suggesting that an instrument may be treated as unidimensional for total score interpretation when a strong general factor exists, even if the content is conceptually multidomain (Donhauser et al., 2023). Nevertheless, given the substantive multidomain nature of the content, future research should examine bifactor or multidimensional IRT models to evaluate the relative contributions of the general and domain-specific factors more explicitly (Embretson & Reise, 2013). Such analyses would provide additional evidence regarding whether the total score primarily reflects a single general construct or a combination of multiple constructs.

This study has several limitations. First, the moderate internal reliability suggests the instrument may benefit from adding or revising items. Second, this study did not include a separate pilot testing phase before large-scale field administration, although the large sample size allowed iterative item analyses within a single dataset. Third, the sample was drawn from a single geographic region, which limited the generalizability of the findings. Fourth, removing several items may have affected the representation of specific indicators within the scientific literacy framework.

Based on these findings, we recommend the following directions for future research. Cross-validation with independent samples is necessary to examine the stability of the item parameters and factor structure. Differential Item Functioning (DIF) analysis is also important to ensure that the instrument functions equivalently across subgroups. Additionally, qualitative approaches, such as cognitive interviews, may help refine items with low discrimination or excessive difficulty. Developing an item bank and applying multidimensional modelling approaches are further strategic steps to enhance the instrument's psychometric quality and applicability.

Conclusion

This study resulted in the development of the HSP-SLT, an instrument that demonstrates stable measurement characteristics based on CTT and IRT analyses. The iterative calibration process enabled the identification and refinement of malfunctioning items, resulting in a final model that satisfied key measurement assumptions, produced interpretable item parameters, and demonstrated a satisfactory overall model fit. The TIF indicated that the instrument performed optimally at moderate to moderately high ability levels, suggesting that ability estimates within this range were obtained with greater precision. The findings contribute both methodologically and practically to the development of scientific literacy assessments in physics education. Methodologically, this study shows that using CTT and IRT is an

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

effective way to produce more precise, validated instruments. From a practical standpoint, the instrument may assist educators in identifying students' scientific literacy at higher conceptual levels.

Acknowledgement

This research was a collaboration between Universitas Sulawesi Barat and Universitas Negeri Makassar, one of the Operational Assistance Programs for Higher Education for the National Competitive Research Program in 2022. We express our deepest gratitude to the Ministry of Education, Culture, Research, and Technology (KEMENDIKBUDRISTEK) for allowing the authors to complete this research.

Conflict of Interest

The authors declare no conflicts of interest.

Authors Contribution

Conceptualization: N.N., M.M., and N.A.L.; methodology: N.N., M.M., and N.A.L.; validation: K.A. and H.A.; formal analysis: N.N., M.M., and N.A.L.; data curation: N.N., M.M., and N.A.L.; writing-original draft preparation: N.N.; writing–review and editing: N.N., M.M., and N.A.L.; supervision: K.A. and H.A. All authors have read and agreed to the published version of this manuscript.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During preparation of this manuscript, the corresponding author used AI-assisted language tools to check English grammar and improve the article's readability. After using these tools, the author reviewed and edited the content as needed and took full responsibility for the publication.

References

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, *33*(9), 1289–1312. https://doi.org/10.1080/09500693.2010.512369

Afnan, R., Munasir, M., Budiyanto, M., & Aulia, M. I. R. (2023). The Role of Scientific Literacy Instruments For Measuring Science Problem Solving Ability. *IJORER : International Journal of Recent Educational Research*, *4*(1), 45–58. https://doi.org/10.46245/ijorer.v4i1.271

Ahorsu, D. K., Lin, C.-Y., Imani, V., Saffari, M., Griffiths, M. D., & Pakpour, A. H. (2022). The Fear of COVID-19 Scale: Development and Initial Validation. *International Journal of Mental Health and Addiction*, *20*(3), 1537–1545. https://doi.org/10.1007/s11469-020-00270-8

Algiranto, A. (2022). Pengembang Instrumen Tes Berbasis Literasi Sains Materi Fluida Statis Kelas XI SMA. *Silampari Jurnal Pendidikan Ilmu Fisika*, *4*(1), 47–58.

Arohman, M., Saefudin, & Priyandoko, D. (2016). Kemampuan Literasi Sains Pada Pembelajaran Ekosistem. *Jurnal FMIPA Universitas Pendidikan Indonesia ISSN : 2528-5742*, *13*(1), 90–92.

Ayanwale, M. A., Amusa, J. O., Oladejo, A. I., & Ayedun, F. (2024). Multidimensional item Response Theory Calibration of Dichotomous Response Structure Using R Language for Statistical Computing. *Interchange*, *55*(1), 137–157. https://doi.org/10.1007/s10780-024-09517-y

Baker, F. B. (2001). *The Basics of Item Response Theory* (C. Boston & L. Rudner (eds.); Second Edi). ERIC Clearinghouse on Assessment and Evaluation.

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

and Item Development. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(2), 142. https://doi.org/10.11591/ijere.v7i2.12900

Bichi, A., Talib, R., Atan, A., Ibrahim, H., & Yusof, S. (2019). Validation of a developed university placement test using classical test theory and Rasch measurement approach. *International Journal of ADVANCED AND APPLIED SCIENCES*, *6*, 22–29. https://doi.org/10.21833/ijaas.2019.06.004

Bijlsma, H., van der Lans, R., Mainhard, T., & den Brok, P. (2021). Psychometric Perspectives: CCT, IRT. *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers*.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bock, R. D., & Gibbons, R. D. (2021). Item Response Theory. In *The SAGE Handbook of Quantitative Methods in Psychology*. John Wiley & Sons, Inc. https://doi.org/10.4135/9780857020994.n7

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q(3): Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

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

DeMars, C. (2010). *Item Response Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

DeMars, C. E. (2018). Classical Test Theory and Item Response Theory. In *The Wiley Handbook of Psychometric Testing* (pp. 49–73). https://doi.org/https://doi.org/10.1002/9781118489772.ch2

Denieffe, S. (2020). Commentary: Purposive sampling: complex or simple? Research case examples. *Journal of Research in Nursing : JRN*, *25*(8), 662–663. https://doi.org/10.1177/1744987120928156

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

DetikNews. (2020, November 26). *Berburu ikan dengan senapan angin*. detikcom. https://news.detik.com/foto-news/d-5270654/berburu-ikan-dengan-senapan-angin?page=3

Development, O. for E. C. and. (2023). *PISA 2022 assessment and analytical framework*. OECD publishing.

Donhauser, T., Apfelbacher, C., Kann, G., Masur, C., Kamudoni, P., Salek, S., Abels, C., & Gabes, M. (2023). Hyperhidrosis quality of life index (HidroQoL): further validation by applying classical test theory and item response theory using data from a phase III clinical trial. *Journal of Patient-Reported Outcomes*, *7*(1), 55. https://doi.org/10.1186/s41687-023-00596-6

Doran, R. L. (1980). *Basic Measurement and Evaluation of Science Instruction*. National Science Teachers Association.

Eaton, P., Johnson, K., Frank, B., & Willoughby, S. (2019). Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, *15*(1), 10102. https://doi.org/10.1103/PhysRevPhysEducRes.15.010102

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

El-Hamamsy, Laila, Zapata-Cáceres, María, Barroso, Estefanía Martín, Mondada, Francesco, Zufferey, Jessica Dehler, & Bruno, Barbara. (2022). The Competent Computational Thinking Test: Development and Validation of an Unplugged Computational Thinking Test for Upper Primary School. *Journal of Educational Computing Research*, *60*(7), 1818–1866. https://doi.org/10.1177/07356331221081753

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.

Finch, W. H., & Jeffers, H. (2016). A Q(3)-Based Permutation Test for Assessing Local Independence. *Applied Psychological Measurement*, *40*(2), 157–160. https://doi.org/10.1177/0146621615622635

Fives, H., Huebner, W., Birnbaum, A. S., & Nicolich, M. (2014). Developing a Measure of Scientific Literacy for Middle School Students. *Science Education*, *98*(4), 549–580. https://doi.org/10.1002/sce.21115

Goldmann, P., Schäfer, B., & Altendorfer, C. (2025). Development and validation of the digital mindset scale. *Journal of Business Research*, *199*, 115554. https://doi.org/https://doi.org/10.1016/j.jbusres.2025.115554

Gormally, C., Brickman, P., & Lut, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education*, *11*(4), 364–377. https://doi.org/10.1187/cbe.12-03-0026

Hartono, W., Hadi, S., & Rosnawati, R. (2024). Optimizing Item Construction in Diagnostic Mathematics Test. *TEM Journal*, 1106–1115. https://doi.org/10.18421/TEM132-25

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*(9 Suppl), II28-42. https://doi.org/10.1097/00005650-200009002-00007

Huang, F., Ye Han, X., Chen, S.-L., Guo, Y. F., Wang, A., & Zhang, Q. (2020). Psychometric Testing of the Chinese Simple Version of the Simulation Learning Effectiveness Inventory: Classical Theory Test and Item Response Theory. *Frontiers in Psychology*, *Volume 11-2020*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.00032

Indrawati, M. D., & Sunarti, T. (2018). Pengembangan Instrumen Penilaian Literasi Sains Fisika Peserta Didik pada Bahasan Gelombang Bunyi di SMA Negeri 1 Gedangan Sidoarjo. *Jurnal Inovasi Pendidikan Fisika (JIPF)*, *07*(01), 14–20.

Istiyono, E. (2020). Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern. *Yogyakarta: UNY Press. L, I.(2019). Evaluasi Dalam Proses Pembelajaran. Jurnal Manajemen Pendidikan Islam*, *9*, 478–492.

Istyadji, M., & Sauqina. (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. https://doi.org/10.36681/tused.2023.016

Kalkbrenner, M. T. (2021). A Practical Guide to Instrument Development and Score Validation in the Social Sciences: The MEASURE Approach. *Practical Assessment, Research & Evaluation*, *26*, 1.

Karimah, K., Jufri, A. W., & Ramdani, A. (2021). Improving Students' Scientific Literacy: Learning Studies Using Science Teaching Materials on Circulatory System Materials. *Prisma Sains : Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, *9*(1 SE-Research Articles), 57–63. https://doi.org/10.33394/j-ps.v9i1.3878

Kong, S.-C., & Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. *Computers & Education*, *187*, 104562. https://doi.org/https://doi.org/10.1016/j.compedu.2022.104562

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Ly, C. Y., Rakkapao, S., Nualtong, K., & Sumathakulawat, W. (2021). Classical test theory and rasch analysis of test of understanding of vectors (TUV). *Journal of Physics: Conference Series*, *1719*(1). https://doi.org/10.1088/1742-6596/1719/1/012089

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Noor, M. S. A. M. (2021). Assessing Secondary Students' Scientific Literacy: A Comparative Study of Suburban Schools in England and Malaysia. *Science Education International*, *32*(4), 343–352. https://doi.org/10.33828/sei.v32.i4.9

Nurjanah, S., Iqbal, M., Zafrullah, Z., Mahmud, M. N., Seran, D. S. F., Suardi, I. K., & Arriza, L. (2024). Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *28*(2 SE-Articles), 161–172. https://doi.org/10.21831/pep.v28i2.71542

Nursyamsudin, N., & Jaelani, M. N. G. (2021). *Penguatan literasi dan numerasi di SMA*. Direktorat Sekolah Menengah Atas.

OECD. (2015). PISA 2015 Draft Science Framework. In *Oecd* (Issue March 2015). https://doi.org/10.1177/0022146512469014

OECD. (2019). PISA 2018 Assessment and Analytical Framework, PISA. In *OECD Publishing*.

OECD. (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education. In *OECD Publishing* (Vol. 46, Issue 183). https://doi.org/10.1787/53f23881-en

Oktalia, Y., Susila, A. B., & Iswanto, B. H. (2021). Virtual Test Instruments to Measure Scientific Literacy of High School Students on Work and Energy. *Journal of Physics: Conference Series*, *2019*(1), 12009. https://doi.org/10.1088/1742-6596/2019/1/012009

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/https://doi.org/10.1002/nur.20147

Qian, Z., Yang, Y., Tan, J., Li, Y., Zhou, J., Huang, J., Assanangkornchai, S., & Chen, J. (2025). Psychometric Evaluation of the Chinese Version of the Mental Health System Responsiveness Questionnaire for Psychiatric Outpatients: Classical Test Theory and Item Response Theory Approaches. *Patient Preference and Adherence*, *19*, 729–740. https://doi.org/10.2147/PPA.S503016

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., & Harrison, M. (2019). On the Connections Between Item Response Theory and Classical Test Theory: A Note on True Score Evaluation for Polytomous Items via Item Response Modeling. *Educational and Psychological Measurement*, *79*(6), 1198–1209. https://doi.org/10.1177/0013164417745949

Revelle, W. (2025). *Package ' psych ': Procedures for Psychological, Psychometric, and Personality Research Description* (2.5.3). https://personality-project.org/r/psych/

Rusilowati, A., Kurniawati, L., Nugroho, S. E., & Widiyatmoko, A. (2016). Developing an instrument of scientific literacy asessment on the cycle theme. *International Journal of Environmental and Science Education*, *11*(12), 5718–5727.

Santoso, P. H., Istiyono, E., Haryanto, & Retnawati, H. (2024). Validating light phenomena conceptual assessment through the lens of classical test theory and item response theory frameworks. *Physics*

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Education, *59*(2), 1–19. https://doi.org/10.1088/1361-6552/ad183b

Sartika, D., Lutfin, N. A., Musdar, M., & Nurlina, N. (2024). Kemampuan Calon Guru Fisika Dalam Menyelesaikan Soal Literasi Sains Berbasis Kearifan Lokal. *Jurnal Penelitian Pembelajaran Fisika*, *15*(1), 49–54. https://doi.org/10.26877/jp2f.v15i1.16683

Sartika, D., Nurlina, Mutmainna, Aris, N. A., & Musliana. (2023). The scientific literacy profile of senior high school student based on science competence dimension. *AIP Conference Proceedings*, *2619*. https://doi.org/10.1063/5.0122563

Sartika, D., Nurlina, N., & Mutmainna, M. (2025). Scientific Literacy Assessment for Physics Instruction by IntegrationLocal Wisdom of Mandar Tribe in West Sulawesi, Indonesia. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *10*(2), 263–272. https://doi.org/10.26737/jipf.v10i2.6839

Şentürk, C., & Sari, H. (2018). Investigation of the contribution of differentiated instruction into science literacy. *Qualitative Research in Education*, *7*(2), 197–237. https://doi.org/10.17583/qre.2018.3383

Setiawati, F. A., Amelia, R. N., Sumintono, B., & Purwanta, E. (2023). Study Item Parameters of Classical and Modern Theory of Differential Aptitude Test: Is it Comparable? *European Journal of Educational Research*, *12*(2), 1097–1107. https://doi.org/10.12973/eu-jer.12.2.1097

Shahzadi, I., & Nasreen, A. (2020). Assessing Scientific Literacy Levels among Secondary School Science Students of District Lahore. *Bulletin of Education and Research*, *42*(3), 1–21.

Sharkness, J., & DeAngelo, L. (2011). Measuring Student Involvement: A Comparison of Classical Test Theory and Item Response Theory in the Construction of Scales from Student Surveys. *Research in Higher Education*, *52*(5), 480–507. https://doi.org/10.1007/s11162-010-9202-3

Sheng, Y. (2019). CTT Package in R. *Measurement*, *17*(4), 211–219. https://doi.org/10.1080/15366367.2019.1600839

Shi, H., Ren, Y., Xian, J., Ding, H., Liu, Y., & Wan, C. (2024). Item analysis on the quality of life scale for anxiety disorders QLICD-AD(V2.0) based on classical test theory and item response theory. *Annals of General Psychiatry*, *23*(1), 1–12. https://doi.org/10.1186/s12991-024-00504-2

Shohib, M., Rahayu, Y., Wasis, W., & Hariyono, E. (2021). Scientific Literacy Ability of Junior High School Students on Static Electricity and Electricity in Living Things. *IJORER : International Journal of Recent Educational Research*, *2*, 700–708. https://doi.org/10.46245/ijorer.v2i6.170

Syamsuddin, S. (2023). Implementasi Classic Test dan Item Respon Theory Pada Penilaian Tes Pembelajaran Matematika. *EDUSCOPE: Jurnal Pendidikan, Pembelajaran, Dan Teknologi*, *8*(2 SE-Articles), 28–43. https://doi.org/10.32764/eduscope.v8i2.3488

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thissen, D. (2015). Psychometrics: Item Response Theory. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (Second Edi, Vol. 19). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.42071-4

Tong, L. K., Au, M. L., Zheng, M. R., Liu, Y. B., Fu, G. L., & Li, Y. Y. (2024). Health behavior inventory – short form: a Chinese validation study using classical test theory and item response theory. *BMC Public Health*, *24*(1), 2281. https://doi.org/10.1186/s12889-024-19774-2

Uyigue, A. V., & Orheruata, M. U. (2019). Test Length and Sample Size for Item-Difficulty Parameter Estimation in Item Response Theory. *Journal of Education and Practice*, *10*(30), 72–75.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

https://doi.org/10.7176/jep/10-30-08

Vilca, L. W., Chávez, B. V., Fernández, Y. S., & Caycho-Rodríguez, T. (2022). Spanish Version of the Revised Mental Health Inventory-5 (R-MHI-5): New Psychometric Evidence from the Classical Test Theory (CTT) and the Item Response Theory Perspective (IRT). *Trends in Psychology*, *30*(1), 111–128. https://doi.org/10.1007/s43076-021-00107-w

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & Brand, T. van den. (2025). *Package 'ggplot2': Create Elegant Data Visualisations Using the Grammar of Graphics* (3.5.2). https://ggplot2.tidyverse.org

Wikcham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Package "dplyr": A Grammar of Data Manipulation* (1.1.4).

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Yang, F., Zhao, F., Zheng, Y., & Li, G. (2020). Modification and verification of the Infant–Toddler Meaningful Auditory Integration Scale: a psychometric analysis combining item response theory with classical test theory. *Health and Quality of Life Outcomes*, *18*(1), 367. https://doi.org/10.1186/s12955-020-01620-9

Zhang, C., Zhang, H., Zhao, M., Liu, D., Zhao, Y., & Yao, Y. (2020). Assessment of Geriatric Depression Scale's Applicability in Longevous Persons based on Classical Test and Item Response Theory. *Journal of Affective Disorders*, *274*, 610–616. https://doi.org/https://doi.org/10.1016/j.jad.2020.05.090

Zhang, D., Wang, C., Yuan, T., Li, X., Yang, L., Huang, A., Li, J., Liu, M., Lei, Y., Sun, L., Zhang, J., & Zhang, L. (2023). Psychometric properties of the Coronavirus Anxiety Scale based on Classical Test Theory (CTT) and Item Response Theory (IRT) models among Chinese front-line healthcare workers. *BMC Psychology*, *11*(1). https://doi.org/10.1186/s40359-023-01251-x

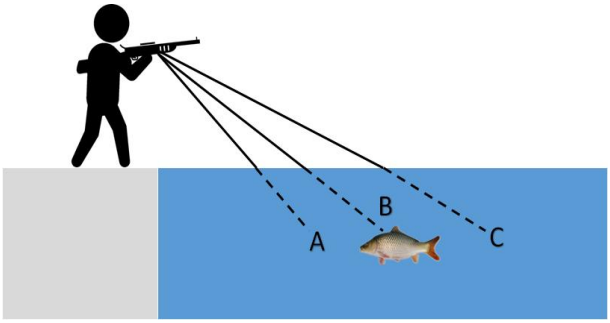
Zhang, L., Liu, X., & Feng, H. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1). https://doi.org/10.1186/s43031-023-00093-2

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Appendix

Instrument (Original Version)

Due to the length of the instrument, which consists of 25 items, only a selection is included in the appendix of this article. Please contact the co-author if you are interested in the complete instrument.

Question (IDES6)
Toni ingin menangkap ikan dengan menggunakan senapan angin. Lokasi yang jelas dari ikan tersebut dapat dilihat pada gambar dan ilustrasi berikut. Karena pembiasan cahaya, lokasi ikan yang teramati adalah bukan lokasi yang sebenarnya.   Agar Toni berhasil mengarahkan senapan anginnya tepat pada ikan, posisi senapan anginnya harus diarahkan … a. Tepat pada titik C b. Tepat pada titik B c. Tepat pada titik A d. Lurus sejajar dengan permukaan air e. Tepat dimana terdapat gelembung air

Competency *(Indicator)*	*Interpreting Data and Evidence Scientifically* **(IDES6)**
Knowledge - System	*Content*
Context	*Personal – Frontiers of Science and Technology*
Cognitive Demand	*Low*

http://journal.uinjkt.ac.id/index.php/jp3i

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Question (EDSE8)

Soni merupakan seorang anak yang bermimpi menjadi atlet angkat besi setelah melihat atlet angkat besi Indonesia Eko Yuli Irawan memenangkan medali emas dalam ajang Asian Games 2018 Jakarta-Palembang. Sejak itu setiap diminta oleh Ibunya mengambil air di sumur belakang rumah, Soni juga berlatih mengangkat 2 ember berisi penuh air dengan kedua tangannya tanpa berpindah posisi selama beberapa detik sebelum membawa 2 ember air tersebut ke Ibunya. Walaupun hal yang dilakukan Soni bukan bentuk latihan profesional, namun Soni menganggapnya sebagai usaha awalnya dalam melatih kekuatan lengannya. Namun Ibu Soni menyampaikan hal menarik kepada Soni bahwa *"Dalam ilmu fisika sebenarnya tidak ada usaha yang dilakukan Soni, walaupun dia setiap hari melakukan latihan tersebut".*

Berdasarkan pernyataan Ibu Soni di atas, maka perkataan Ibu Soni …

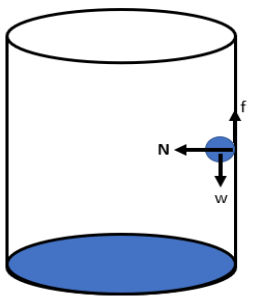
a. Salah, karena selama melakukan latihan tersebut Soni memberikan gaya pada ember
b. Salah, karena Soni merasa lelah saat melakukan latihan tersebut
c. Salah, karena ada energi yang dikeluarkan Soni selama melakukan latihan tersebut
d. Benar, usaha yang dilakukan Soni sama dengan nol karena Soni tidak melakukan perpindahan posisi selama melakukan latihan tersebut
e. Benar, usaha yang dilakukan Soni sama dengan nol karena Soni memberikan gaya pada ember

Competency *(Indicator)*	*Evaluating and Designing Scientific Enquiry* **(EDSE8)**
Knowledge - System	*Content*
Context	*Personal – Frontiers of Science and Technology*
Cognitive Demand	*Low*

Question (EPS1)

Permainan tong stan/tong stand/*wall of death* atau disebut juga roda gila adalah sebuah acara karnival yang menampilkan sebuah silinder kayu berbentuk kerucut atau tabung. Ukuran silindernya umumnya memiliki diamater 20 – 36 ft (6 – 11 m). Saat dilakukan atraksi, para pengemudi sepeda motor atau pengemudi mobil menyetir disepanjang dinding vertikal.





This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Pernyataan manakah yang tepat berdasarkan deksripsi dan gambar yang disajikan di atas? a. Pergerakan kendaraan dalam tong silider dimulai dari permukaan atas silider dengan kecepatan rendah b. Kendaraan dalam tong silider akan tetap bisa berada di dinding walau tanpa kecepatan sekalipun c. Kecepatan gerakan kendaraan dalam tong silinder tidak dipengaruhi oleh massa kendaraan tersebut d. Jumlah kendaraan yang dapat beroperasi dalam tong silinder tidak dipengaruhi oleh diameter silinder e. Gaya berat yang terjadi pada lebih kecil dibanding besarnya gaya gesek pada dinding dan gaya sentripetal kendaraan tersebut.

Competency *(Indicator)*	*Explaining Phenomena Scientifically* (**EPS1**)
Knowledge - System	*Content*
Context	*Local/National – Natural Resources*
Cognitive Demand	*Medium*

Scoring

The scoring system for this study follows a dichotomous approach, where each item is assigned a binary value:

- **Correct response (1)**
- **Incorrect response (0)**

http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Classical Test Theory (CTT)

Classical Test Theory Item Analysis Based on Doran (1980)

Item	Difficulty	Discrimination	Diff._Category	Disc._Category	Category	Decision
EPS1	0.232	0.356	Acceptable	Moderate	Acceptable	Retain
EPS2	0.314	0.104	Acceptable	Low	Needs Revision	Revise
EPS3	0.39	0.385	Acceptable	Moderate	Acceptable	Retain
EPS4	0.378	0.324	Acceptable	Moderate	Acceptable	Retain
EPS5	0.382	0.196	Acceptable	Low	Needs Revision	Revise
EPS6	0.238	0.408	Acceptable	Moderate	Acceptable	Retain
EPS7	0.257	0.402	Acceptable	Moderate	Acceptable	Retain
EPS8	0.244	0.352	Acceptable	Moderate	Acceptable	Retain
EPS9	0.302	0.333	Acceptable	Moderate	Acceptable	Retain
EPS10	0.306	0.255	Acceptable	Moderate	Acceptable	Retain
EPS11	0.238	0.359	Acceptable	Moderate	Acceptable	Retain
EDSE1	0.343	0.355	Acceptable	Moderate	Acceptable	Retain
EDSE2	0.255	0.22	Acceptable	Moderate	Acceptable	Retain
EDSE3	0.224	0.183	Acceptable	Low	Needs Revision	Revise
EDSE4	0.281	0.385	Acceptable	Moderate	Acceptable	Retain
EDSE5	0.224	0.347	Acceptable	Moderate	Acceptable	Retain
EDSE6	0.138	0.014	Difficult	Low	Needs Revision	Revise
EDSE7	0.218	0.292	Acceptable	Moderate	Acceptable	Retain
EDSE8	0.316	0.292	Acceptable	Moderate	Acceptable	Retain
IDES1	0.205	0.291	Acceptable	Moderate	Acceptable	Retain
IDES2	0.185	0.147	Difficult	Low	Needs Revision	Revise
IDES3	0.173	0.155	Difficult	Low	Needs Revision	Revise
IDES4	0.226	0.013	Acceptable	Low	Needs Revision	Revise
IDES5	0.255	0.281	Acceptable	Moderate	Acceptable	Retain
IDES6	0.248	0.419	Acceptable	Moderate	Acceptable	Retain

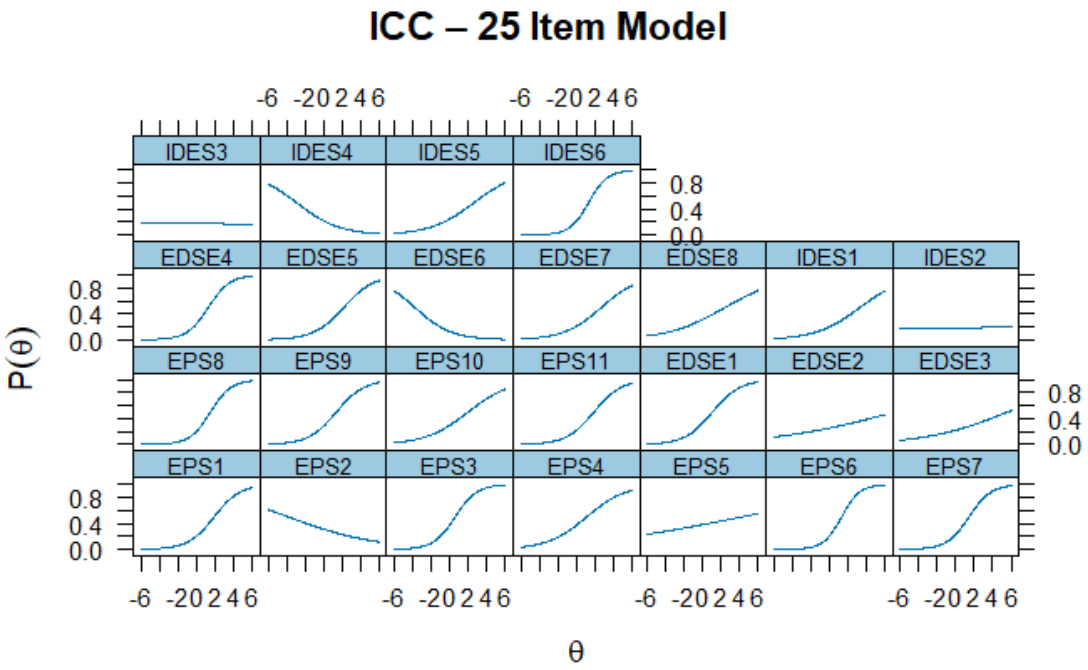
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

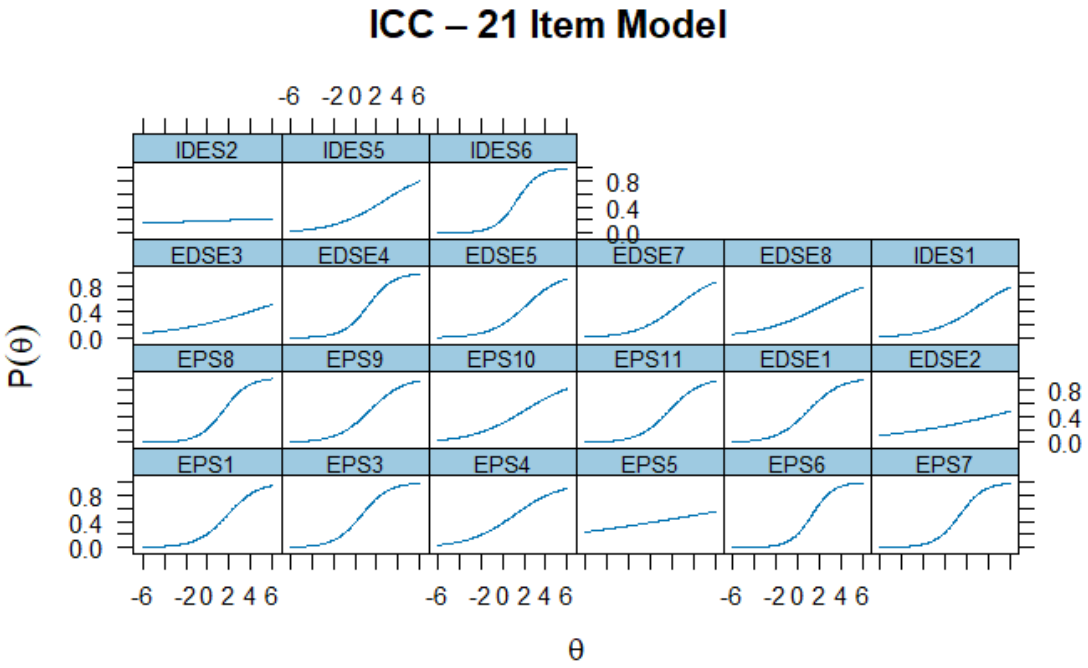
Item Response Theory (IRT)

Comparison of Item Discrimination (a) and Difficulty (b) Parameters Across the 25-, 21-, and 20-Item 2-PL Models

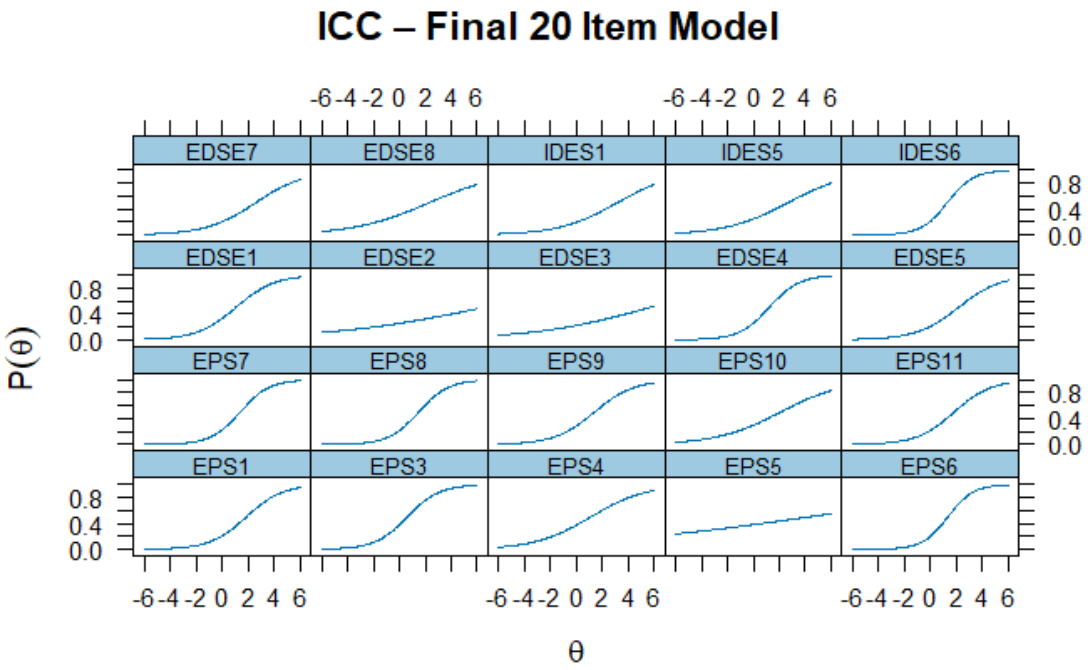
Item	a			b		
	25 items	**21 items**	**20 items**	**25 items**	**21 items**	**20 items**
EPS1	0.71595	0.712641	0.714564	1.835864	1.843296	1.839059
EPS2	-0.2075			-3.80901		
EPS3	0.860851	0.804748	0.806302	0.577463	0.611917	0.610848
EPS4	0.47199	0.468186	0.465883	1.098982	1.107347	1.112414
EPS5	0.115217	0.113673	0.112926	4.185888	4.242369	4.270269
EPS6	1.027619	1.024213	1.026799	1.336991	1.340831	1.338206
EPS7	0.839855	0.890507	0.889271	1.423178	1.358179	1.359596
EPS8	0.898143	0.868679	0.868314	1.442231	1.481362	1.48178
EPS9	0.668306	0.646867	0.647702	1.358511	1.397777	1.396159
EPS10	0.435716	0.403689	0.405983	1.953386	2.097935	2.086791
EPS11	0.676969	0.69214	0.691868	1.872938	1.838028	1.838601
EDSE1	0.664135	0.669321	0.667627	1.056899	1.049898	1.052179
EDSE2	0.154552	0.165043	0.163884	6.961846	6.524125	6.56972
EDSE3	0.225842	0.218347	0.219739	5.55844	5.745155	5.709497
EDSE4	0.860898	0.864151	0.864799	1.236481	1.233034	1.232219
EDSE5	0.61557	0.609332	0.612115	2.167905	2.187496	2.178758
EDSE6	-0.50528			-3.78918		
EDSE7	0.494665	0.519019	0.518881	2.707691	2.592215	2.592818
EDSE8	0.321148	0.338711	0.339184	2.461695	2.339449	2.336318
IDES1	0.41968	0.44553	0.443675	3.352466	3.171413	3.183669
IDES2	0.008933	0.039536		165.859	37.48689	
IDES3	-0.00291			-535.822		
IDES4	-0.43056			-2.97105		
IDES5	0.421843	0.411801	0.41199	2.632165	2.692099	2.690926
IDES6	0.990496	0.997977	0.993095	1.31131	1.304325	1.309098

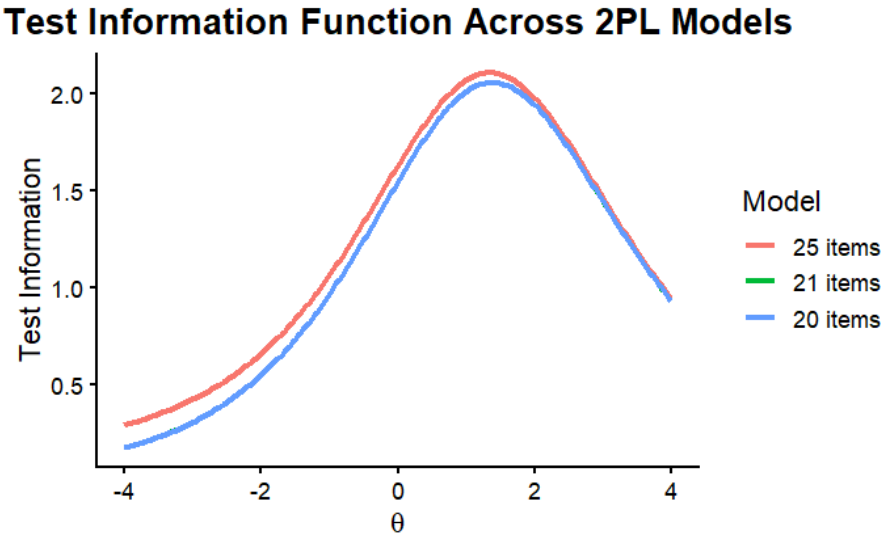
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)





This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)





This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)