Science Education and Development Journal Archives
Vol. 4, No. 1, pp. 1-15, Maret 2026 | e-ISSN: 2988-5337
https://ejournal.imbima.org/index.php/sendja
https://doi.org/10.59923/sendja.v4i1.654

Measuring 21st-century science laboratory competence: development and validation of a contextual assessment tool

Andina Nurul Wahidah[1]*, Inayah Dzil Izzati Hartono[2]

[1]Mathematics Education Study Program, Pontianak State Islamic Intitute, Pontianak, Indonesia
[2]Educational Research and Evaluation Study Program, Yogyakarta State University, Yogyakarta, Indonesia

**correspondence email: andinanurulwahidah@iainptk.ac.id*

Abstract

Mastery of contextual experiment-based science laboratory competencies is essential for strengthening scientific literacy, critical thinking, and data interpretation skills in the 21st century. However, there is a lack of standardized instruments that holistically measure conceptual, procedural, and interpretative dimensions, particularly at the secondary education level. This study aimed to develop and validate an assessment instrument for Contextual Experiment-Based Biology Laboratory Competency using a limited research and development (R&D) design. Instrument development involved theoretical analysis, blueprint construction, expert validation, pilot testing, Exploratory Factor Analysis (EFA), and Rasch model analysis. Item development was grounded in literature review, national curriculum guidelines, and locally relevant experimental contexts integrated into students' learning experiences. Content validation by eight experts using I-CVI and S-CVI yielded high agreement (I-CVI ≥ 0.87; S-CVI = 0.95). Construct validity was examined using Exploratory Factor Analysis (EFA) with Maximum Likelihood extraction and Promax rotation on trial data from 135 Madrasah Aliyah students. The results showed a KMO value of 0.75 and a significant Bartlett's Test ($p < 0.001$). Three major factors emerged, explaining 43.0% of the total variance and aligning with the initial construct framework. Further calibration using the Rasch Model demonstrated good item fit, high reliability (0.86), and well-distributed item logits. The instrument proved valid and reliable as a diagnostic assessment tool based on contextual experimentation. The findings support the implementation of authentic, context-based assessment and recommend follow-up CFA analysis and practical application in biology laboratory instruction.

Keywords: Laboratory Competency; Contextual Experiment; Instrument Validation; EFA; Rasch Model

Introduction

The development of laboratory competence in science education is a critical priority in preparing students for the challenges of the twenty-first century. Laboratory skills reflect not only mastery of scientific content but also indicators of scientific literacy, encompassing conceptual understanding, procedural skills, and interpretative abilities (Adamczyk A., 2025; Mauldin, 2025). Integrating problem-based learning (PBL) into laboratory courses enhances experimental design, data analysis, collaborative problem-solving, and adaptability, providing students with authentic scientific experiences aligned with global education standards (Adamczyk A., 2025). Similarly, inquiry-driven learning approaches increase student engagement across cognitive, emotional, and social dimensions, strengthening curiosity, critical thinking, and motivation in science learning (Redman C., 2021; Wu H. K., 2020).

Contextualized laboratory learning further bridges theoretical knowledge with real-world applications, particularly in contexts where physical laboratory access is limited. For instance, home-based, smartphone-assisted, and asynchronous online experiments enable meaningful investigations despite constraints such as remote learning or limited laboratory resources, promoting deeper engagement and practical skill development (Wu X., 2024; Zhu M., 2021). Extended Remote Laboratories (XRLs) and virtual reality (VR)-supported laboratories offer immersive and interactive experiences that complement traditional settings, enabling students to simulate complex experiments and develop procedural competence even in resource-constrained environments (da Silva R., 2023; Evgenia P. E., 2021). Moreover, structured asynchronous online laboratories have demonstrated learning outcomes comparable to, or even exceeding, conventional in-person laboratories, particularly when combined with formative feedback and systematic performance monitoring (Dao T., 2024; Faulconer E. K., 2018).

Despite these advances, laboratory competence assessment often remains fragmented, focusing either on technical execution or conceptual understanding rather than providing an integrated measure of scientific competence. Recent empirical studies reveal important limitations in existing laboratory assessment instruments. For example, several instruments primarily emphasize procedural performance while providing limited evidence of construct validity and multidimensional measurement of scientific competence (Kota M., 2024). Other studies report that laboratory evaluation tools frequently rely on teacher judgment without adequate control for rater severity, item difficulty, or measurement invariance, potentially reducing the objectivity and reliability of assessment results (Higde E., 2024).

In addition, many currently available instruments were developed for conventional laboratory settings and do not sufficiently incorporate contextual and authentic problem-solving scenarios relevant to twenty-first-century science learning environments. Recent findings also indicate that affective and interpretative aspects of laboratory competence are often underrepresented compared to procedural indicators, despite their importance for scientific literacy and inquiry-based learning outcomes (Higde E., 2024; Kota M., 2024). These limitations highlight the urgent need for a comprehensive and psychometrically robust assessment instrument capable of measuring cognitive, procedural, and interpretative dimensions simultaneously while ensuring objective measurement through advanced analytical approaches such as Rasch modeling. Frameworks such as the Many-Facet Rasch Model offer robust evaluation by controlling for rater bias and item difficulty, producing objective measures of student performance (Higde E., 2024).

From a theoretical perspective, laboratory competence comprises three interrelated dimensions:

1. Conceptual competence – mastery of scientific principles (Adamczyk A., 2025).
2. Procedural competence – ability to design and execute experiments systematically (da Silva R., 2023; Mauldin, 2025).
3. Interpretative competence – capacity to analyze data, draw conclusions, and evaluate results scientifically (Evgenia Paxinou E., 2021; Wu X., 2024).

These dimensions align with the principles of active and inquiry-based learning, where students engage in experiments that integrate cognitive, technical, and affective skills. Research indicates that combining PBL, formative feedback, and technology-supported experimentation not only enhances engagement but also strengthens competency outcomes in laboratory settings (Adamczyk A., 2025; Dao T., 2024; Mauldin, 2025; Redman C., 2021).

Given these considerations, developing a contextualized and validated instrument for assessing biology laboratory competence is imperative. Such an instrument should integrate the three dimensions of competence, incorporate authentic and context-relevant scenarios, and utilize robust psychometric approaches, including Exploratory Factor Analysis and Rasch modeling, to ensure construct validity, reliability, and interpretability (Higde E., 2024; Kota M., 2024). The resulting assessment tool can guide educators in improving laboratory instruction, supporting students' scientific literacy, and fostering the application of biological knowledge in real-world contexts.

Method

Research Design

This study adopted a limited research and development (R&D) design to construct and validate a contextual assessment tool for measuring 21st-century science laboratory competence. The design emphasizes the early phases of instrument development (item construction, expert validation, and empirical validation of construct validity and reliability) drawing on established frameworks for instrument development and psychometric validation (Boone et al., 2014; Sumintono & Widhiarso, 2015; DeVellis, 2016). The development process consisted of six stages:

1. Theoretical analysis of science laboratory competence aligned with 21st-century skills frameworks (OECD, 2018; Voogt & Roblin, 2012);

2. Development of an item blueprint reflecting three competence dimensions: conceptual, procedural, and interpretative;
3. Expert validation using the Item- and Scale-level Content Validity Index (I-CVI, S-CVI);
4. Pilot testing with students engaged in contextualized biology laboratory practices;
5. Exploratory Factor Analysis (EFA) to examine the latent structure of competence;
6. Rasch model analysis to refine item functioning and ensure measurement robustness.

This dual approach, combining classical psychometrics through Exploratory Factor Analysis (EFA) and modern measurement theory through Rasch modelling, was employed to capture both the dimensionality and measurement invariance required for robust educational assessment in science (Bond & Fox, 2015; N. K. Fidan, 2020). EFA was used to identify the underlying factor structure of the instrument and to examine the relationships among laboratory competence indicators empirically. Meanwhile, Rasch model analysis was applied to evaluate item functioning, person ability estimation, item fit, and reliability, thereby ensuring that the instrument produced objective and stable measurements across varying levels of student competence.

The instrument was theoretically grounded in models of 21st-century competencies and contextual science learning (Bybee, 2010; Pellegrino J. W., 2012). These theoretical perspectives emphasize the integration of conceptual understanding, practical inquiry skills, and scientific reasoning in laboratory learning. Accordingly, the instrument was designed to reflect authentic laboratory practices relevant to contemporary science education. Three interrelated dimensions of laboratory competence were operationalized:

1. Conceptual competence: understanding principles of biology, theoretical foundations, and the rationale of laboratory inquiry;
2. Procedural competence: practical ability to design, conduct, and control variables in experiments;
3. Interpretative competence: skills in analyzing data, drawing valid conclusions, and evaluating the credibility of results.

From these dimensions, a blueprint was created to guide the development of assessment items systematically. Based on the blueprint, 35 multiple-choice items with four response options were generated to represent the identified competence dimensions comprehensively. The use of multiple-choice items was intended to facilitate objective scoring and efficient administration in classroom assessment contexts.

Items were embedded in local and curriculum-relevant biological contexts such as fermentation, water pollution, photosynthesis, and digestion. The contextualization process aimed to ensure that students could relate scientific concepts to authentic real-world situations encountered in everyday life. This approach also strengthened the cognitive and contextual authenticity of the assessment instrument (Ananiadou K., 2009; Lau W. W. F., 2015).

Content Validation

Content validity was established through expert review involving five validators: two university faculty members in biology education, one experienced biology teacher, and two postgraduate students specializing in curriculum and educational evaluation. Expert judgment is considered a fundamental procedure in instrument development because it ensures that items adequately represent the intended construct domain before empirical testing is conducted (DeVellis, 2016; Polit & Beck, 2006). The inclusion of both academic experts and practitioners also strengthened the ecological and curricular relevance of the developed instrument within authentic biology learning contexts.

A 4-point scale ranging from 1 (not relevant) to 4 (highly relevant) was used to evaluate each item. The validation process covered four aspects: (1) content relevance, (2) clarity, (3) language appropriateness, and (4) technical feasibility. The use of a four-point relevance scale is recommended in content validation studies because it minimizes neutral responses and improves agreement precision among experts (Lynn, 1986; Polit & Beck, 2006).

The Item-level Content Validity Index (I-CVI) and Scale-level Content Validity Index (S-CVI) were computed following the procedures proposed by Polit and Beck (2006). Threshold values of I-CVI ≥ 0.78 and S-CVI ≥ 0.90 were considered acceptable indicators of content validity, particularly when involving five or more experts (Polit D. F., 2007). Items that did not meet these criteria were revised before pilot testing to improve conceptual clarity, language precision, and curricular alignment, consistent with best practices in educational instrument development (DeVellis, 2016).

Participants and Data Collection

The pilot study was conducted with 135 tenth-grade students from a Madrasah Aliyah Negeri (Islamic Senior High School) in Pontianak, Indonesia. The participants were selected because they had prior experience with context-based laboratory learning activities in biology classes. Pilot studies involving samples between 100 and 200 participants are generally considered adequate for preliminary factor analysis and psychometric evaluation in educational research (Comrey A. L., 1992; Hair et al., 2019).

A purposive sampling strategy was employed to select participants who matched the objectives of the research. This sampling approach is widely used in developmental and validation studies because it allows researchers to recruit respondents with relevant learning experiences and contextual familiarity (Creswell & Creswell, 2018). The sampling procedure ensured that students possessed sufficient exposure to laboratory inquiry and contextual biology instruction, enabling more accurate evaluation of the instrument's applicability.

Participation in the study was voluntary, and both students and school administrators were informed about the objectives and procedures of the research before data collection began. Ethical considerations were strictly maintained throughout the study, including informed consent, anonymity protection, and institutional permission. These procedures align with international ethical standards for educational research involving human participants (Cohen L., 2018).

Data Analysis

Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA) was conducted as an initial procedure to empirically examine the dimensionality of the instrument. EFA is commonly applied in the early stages of instrument development to identify latent constructs and evaluate the empirical structure underlying item responses (Costello & Osborne, 2005; Fabrigar et al., 1999). This analysis was intended to determine whether the developed items adequately reflected the theoretical dimensions of laboratory competence.

Sampling adequacy was evaluated using the Kaiser-Meyer-Olkin (KMO) index with a minimum acceptable threshold of 0.60 and Bartlett's Test of Sphericity with a significance level of $p < 0.05$. These procedures are recommended to ensure that inter-item correlations are sufficiently strong for factor analysis (Field, 2018; Hair et al., 2019). Adequate KMO values and significant Bartlett's test results indicate that the correlation matrix is appropriate for extracting latent factors.

Maximum Likelihood extraction with Promax rotation was employed to accommodate the possibility of correlated factors among competence dimensions. Oblique rotation methods such as Promax are recommended in social science research because psychological and educational constructs are often theoretically interrelated (Tabachnick & Fidell, 2019). Factor retention decisions were guided by eigenvalues greater than 1, scree plot inspection, and item factor loadings of at least 0.40, which are widely accepted criteria in psychometric research (Hair et al., 2019).

Items failing to meet the loading criteria were considered for revision or removal to improve construct validity. This refinement process is consistent with recommendations that weak or cross-loading items should be modified to strengthen factor interpretability and measurement accuracy (Costello & Osborne, 2005). Consequently, EFA served as an essential procedure for establishing the internal structure validity of the instrument.

Rasch Model Analysis

Following EFA, Rasch modeling was conducted using the eRm package in RStudio to further evaluate item functioning and measurement quality. Rasch analysis is widely recognized as an advanced psychometric approach because it transforms ordinal raw scores into interval-level measures and provides sample-independent parameter estimation (Bond & Fox, 2015; Boone et al., 2014). The application of Rasch modeling therefore strengthened the precision and objectivity of the developed assessment instrument. Several evaluation criteria were applied during the Rasch analysis process. Evaluation criteria included:

1. Item fit: acceptable range $0.5 < \text{MnSq} < 1.5$;
2. Item difficulty: logits between -2.0 and $+2.0$;
3. Reliability and separation: person/item reliability ≥ 0.80;
4. Wright Map: to examine alignment of item difficulty with student ability distribution.

In addition, person and item reliability coefficients of at least 0.80 were used as indicators of acceptable measurement consistency. Reliability and separation indices in Rasch analysis provide information regarding the instrument's ability to distinguish respondents based on ability levels and classify item difficulty hierarchies accurately (Boone et al., 2014). High reliability values indicate stable and dependable measurement performance across respondents and items.

A Wright Map was generated to examine the alignment between item difficulty distribution and student ability levels. This analysis is important because it visually evaluates whether the instrument appropriately targets participant ability ranges and identifies potential gaps in item distribution (Bond & Fox, 2015). Items showing substantial misfit were reviewed for revision or removal to improve the psychometric quality and interpretability of the instrument.

Software and Ethical Statement

All statistical analyses were conducted in RStudio using the psych, GPArotation, and eRm packages. These packages are widely utilized in psychometric and educational measurement studies because they support comprehensive procedures for factor analysis, rotation techniques, and Rasch modeling (Mair P., 2007; Revelle, 2023). The use of open-source statistical software also promotes transparency, reproducibility, and accessibility in scientific research.

Ethical approval for the study was obtained at both the institutional and school levels prior to data collection. Participation was voluntary, and informed consent was secured from all participants before inclusion in the study. These procedures are essential components of ethical educational research and are intended to protect participant autonomy, privacy, and well-being (Cohen L., 2018).

Given the non-interventional nature of this educational research, formal ethics board approval was not required under local regulations. Nevertheless, all principles of research integrity, confidentiality, participant anonymity, and voluntary participation were consistently maintained throughout the study. Adherence to these ethical standards ensured that the research process remained aligned with accepted international principles for human-subject research.

Results and Discussion

This section presents the empirical evidence from the pilot validation of the contextual assessment tool for 21st-century science laboratory competence. We organize the results in three linked stages: (1) content-validity evidence from an expert review panel, (2) empirical dimensionality exploration using Exploratory Factor Analysis (EFA), and (3) item calibration and fit diagnostics using Rasch modeling. For each stage we first report the quantitative outcomes and then discuss their meaning in relation to the theoretical construct (conceptual, procedural, interpretative competence) and prior work on laboratory assessment, technology-mediated lab instruction, and modern psychometric practice (Adamczyk A., 2025; DeVellis, 2016; Higde E., 2024; Kota M., 2024; Polit & Beck, 2006). Finally, we integrate (cross-sectionally) EFA and Rasch findings to evaluate construct coherence and to identify items for revision or retention. The discussion emphasizes

implications for classroom assessment, curriculum alignment with 21st-century skills (OECD, 2019), and future uses of the instrument for diagnostic and formative feedback (Dao T., 2024; Mauldin, 2025).

Content validity: expert review

Content validity is a foundational step in instrument development to ensure items represent the intended domain comprehensively and meaningfully (DeVellis, 2016; Polit & Beck, 2006). In this study, ten experts were invited to review the initial item pool; eight completed the review. The expert panel comprised a balanced mix of disciplinary, pedagogical, and measurement perspectives: one professor in research & evaluation, two experienced secondary-level biology teachers, two postgraduate students in biology education, and three postgraduate students in measurement and evaluation. This mix follows best practice recommendations to include both content specialists and assessment methodologists so that items are judged for scientific accuracy, curricular alignment, and measurement quality (Fidan, 2020; Waltz C. F., 2010).

Each reviewer evaluated 30 items across four criteria: (a) content relevance to the target indicator, (b) item construction/format, (c) language clarity, and (d) technical appropriateness. A four-point rating scale was used (1 = not appropriate to 4 = highly appropriate); ratings of 3 or 4 were treated as evidence of acceptable content (Polit & Beck, 2006). Item-level Content Validity Index (I-CVI) and the scale-level average S-CVI/Ave were computed following established procedures (Polit & Beck, 2006; Zamanzadeh V., 2015). With eight completing experts, the conventional minimum threshold for acceptable I-CVI was set at 0.78 (Polit & Beck, 2006).

The expert review yielded strong content-validity evidence. I-CVI values ranged from 0.875 to 1.00 across items, indicating high agreement that items mapped to their intended indicators. The overall S-CVI/Ave was 0.94, exceeding common benchmarks for excellent scale-level content validity (Waltz et al., 2010). No item fell below the 0.78 threshold. Several items achieved unanimous endorsement (I-CVI = 1.00), which signals clear consensus on relevance and clarity. Reviewers also supplied minor editorial suggestions (wording standardization, simplification of distractors, clarifying stem phrasing); these were incorporated into the revised item set prior to field testing. Figure 1 illustrates the distribution of I-CVI values across items (red line = 0.78 threshold).

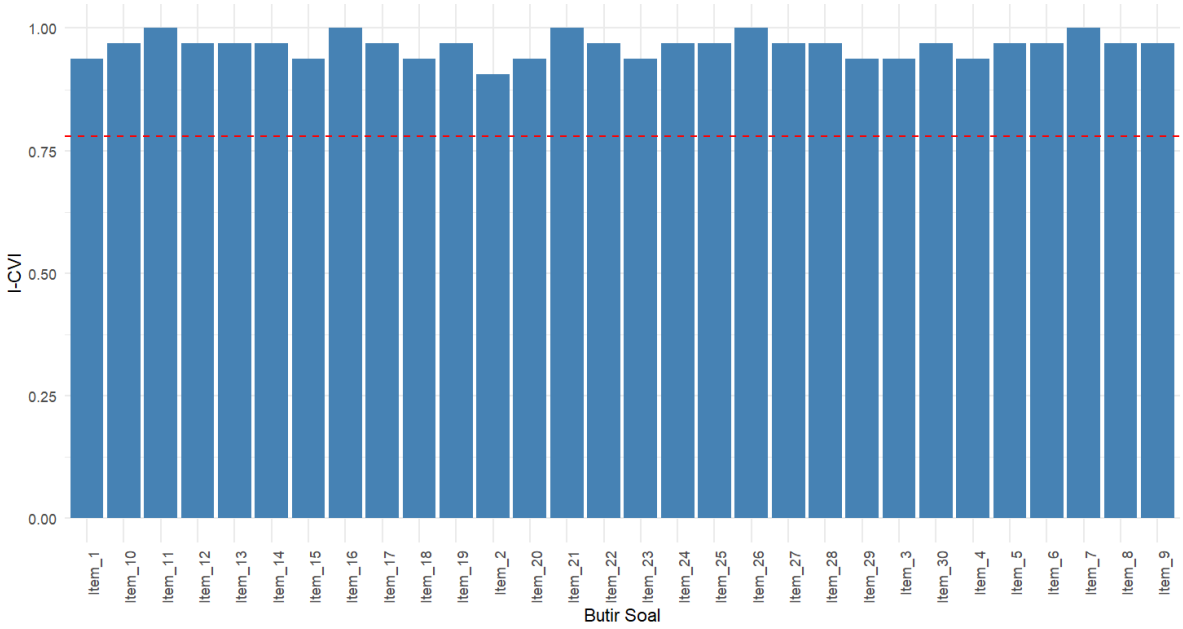

Figure 1. Distribution of I-CVI values per item.
Note. Red horizontal line denotes the minimum acceptable I-CVI (0.78).

The high content-validity indices provide strong initial evidence that the items adequately cover the three hypothesized competence dimensions (conceptual, procedural, interpretative). This expert endorsement aligns with modern calls to ground laboratory assessment in authentic, contextually relevant tasks and to involve practitioners during item construction (Adamczyk A., 2025; OECD, 2018). It also mirrors findings from instrument development studies in laboratory education that emphasize multi-stakeholder validation to achieve both curricular fidelity and assessment utility (Higde E., 2024; Kota M., 2024). Importantly, the reviewers' minor editorial comments focused on wording and response options rather than on construct mismatch, which strengthens confidence that the item pool is conceptually coherent and ready for empirical structural validation (EFA) and Rasch calibration.

The strong content validity justifies proceeding to empirical analyses that probe latent structure and item functioning. In the next sections we report how EFA clarified the emergent factor structure and how Rasch diagnostics informed item retention, local dependence, and measurement invariance, critical steps to produce an assessment that is psychometrically defensible and practically useful for formative feedback in 21st-century laboratory instruction (Dao T., 2024; Mauldin, 2025).

Exploratory factor structure (EFA)

Building on the robust content-validity evidence from expert review, we examined whether learners' response patterns empirically reproduce the theoretically proposed three-domain construct (conceptual, procedural, interpretative). Exploratory Factor Analysis (EFA) was selected because it allows discovery of the instrument's latent structure without imposing a confirmatory model a priori (Hair et al., 2019). EFA therefore functions as the critical bridge from expert judgement to item-level psychometrics, revealing how items cluster in practice and informing subsequent Rasch calibration. Table 1 presents the key fit statistics of the three-factor EFA model.

Table 1. Summary of EFA Statistics – Three-Factor Model

EFA Statistics	Value
Number of Respondents (n)	135
Number of Items	30
Extracted Factors	3
KMO (Overall MSA)	0.75
Bartlett's Test of Sphericity	$\chi^2 = 877.03$; $p < 0.001$
RMSR	0.06
RMSEA	0.00 (CI: 0 – 0.026)
TLI	1.042
BIC	–1373.53
χ^2 (fit model)	333.51; df = 348; p = 0.70
Total Variance Explained	23.0%
Factor 1	12.0%
Factor 2	8.0%
Factor 3	3.0%
Mean Item Complexity	1.05

Prior to extraction we verified factorability: the overall KMO (MSA) = 0.75, indicating adequate sampling adequacy, and Bartlett's Test of Sphericity was highly significant ($\chi^2 = 877.03$, $p < .001$), supporting the suitability of the correlation matrix for factor analysis. Although the initial Kaiser criterion produced several components (10 eigenvalues > 1), the scree plot showed a clear elbow that together with theoretical parsimony supported retention of a three-factor solution aligned with the instrument's design and prevailing frameworks for laboratory competence (Millar, 2004; OECD., 2023).

Analytic choices and overall fit. We used minimum-residual (minres) extraction with Promax rotation to allow correlated factors where theoretically appropriate (e.g., conceptual and procedural skills). Fit and diagnostic indices for the three-factor model were strong: RMSR = 0.06; RMSEA = 0.00 (90% CI: 0.00–0.026); TLI = 1.042; χ^2 (fit) = 333.51, df = 348, p = 0.70. Mean item complexity (~1.05) indicates that most items dominantly load on a single factor, a favourable property for constructing coherent subscales and for subsequent Rasch calibration (Reise S. P., 2000).

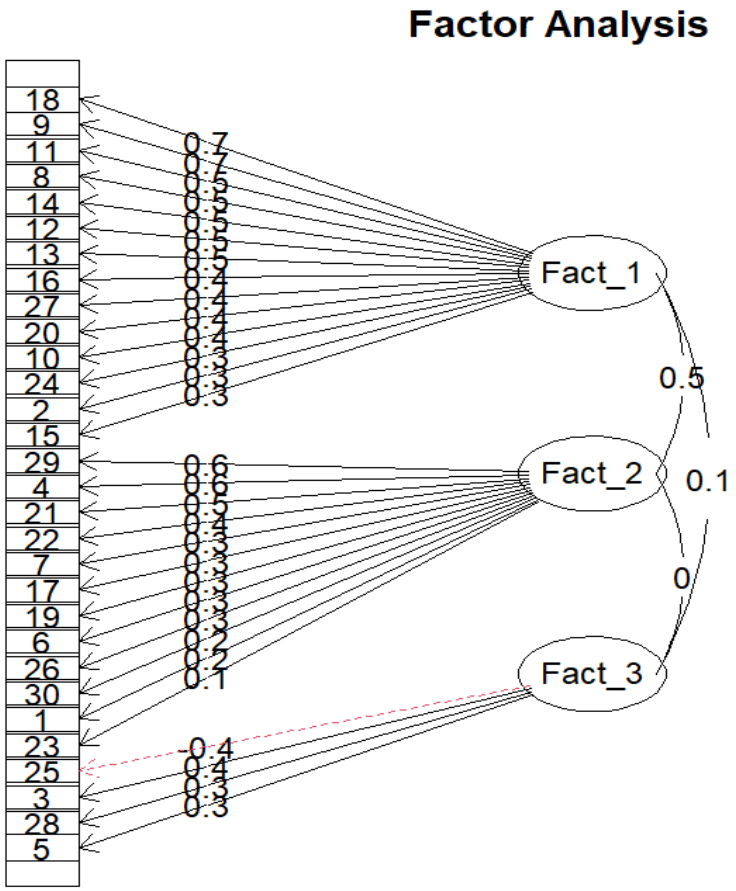

Figure 2. Factor loading diagram of the three-factor EFA model.

Empirical factor composition and loadings. Using a substantive-loading cutoff of |.40|, items grouped as follows:

1. Factor 1 — Conceptual competence (F1): Items 18, 9, 11, 8, 14, 12, 13, 16, 27, 20, 10, 24, 2, 15. These probe conceptual understanding and rationale behind experimental designs.
2. Factor 2 — Procedural competence (F2): Items 29, 4, 21, 22, 7, 17, 19, 6, 26, 30, 1, 23. These address experimental design choices, variable control, and hands-on technique.
3. Factor 3 — Interpretative competence (F3): Items 25*, 3, 28, 5. These assess analysis, interpretation, and evaluation of experimental data.
4. (*Item 25 loaded negatively [−.45] — see note below.)

Although the three factors collectively explained 23% of total variance (F1 = 12%; F2 = 8%; F3 = 3%), the pattern of stable, substantive loadings and strong fit indices supports the interpretability and practical utility of the three-scale solution consistent with similar multi-domain instruments in science-lab research (Faulconer E. K., 2018; Paxinou E., 2021).

Table 2. Factor intercorrelations

Factor	Korelasi Faktor-Skor	Min. Score Correlation	Inter-Factor Correlations		
			F1	F2	F3
F1	0,92	0,69	1,00	0,54	0,12
F2	0,89	0,57	0,54	1,00	0,03
F3	0,73	0,08	0,12	0,03	1,00

Note: Values on the diagonal represent the correlation of each factor with itself (1.00).

Inter-factor relations and construct distinctiveness. Inter-factor correlations were moderate to low: F1–F2 = .54 (moderate correlation), F1–F3 = .12, F2–F3 = .03 (near zero). This pattern indicates that conceptual and procedural competences are related congruent with the view that conceptual knowledge supports procedural skill while interpretative competence behaves relatively independently, reflecting its higher-order analytic nature (Kota M., 2024; Wu H. K., 2020). Factor-score reliabilities were high (F1 = .92; F2 = .89; F3 = .73), indicating that extracted scores represent their intended constructs well.

On the negative loading (Item 25). Item 25 exhibited a substantive negative loading (−.45) on the interpretative factor. A content review revealed reverse-wording and negative polarity that respondents likely processed in the opposite direction to the positively keyed interpretative items (Brown, 2009). We therefore flagged Item 25 for rewording (to positive polarity) or for scoring as a reverse item; it was retained provisionally to allow evaluation of its behaviour under Rasch analysis and potential DIF checks (Higde E., 2024).

Practical implications. The EFA supports scoring the instrument as three subscales (Conceptual, Procedural, Interpretative). Teachers and researchers can use subscale profiles for diagnostic purposes (e.g., targeting procedural coaching vs. interpretation scaffolds), and curriculum designers can align interventions (PBL, formative feedback, smartphone/XR-supported lab tasks) with the domain-specific needs indicated by the instrument (Adamczyk A., 2025; Dao T., 2024).

Link to item-level psychometrics. While EFA establishes an interpretable latent structure, it does not test item-level functioning such as invariance, fit, or person–item targeting. Accordingly, we proceeded to Rasch calibration (Section 3.3) to (a) evaluate fit statistics (infit/outfit), (b) estimate item difficulty (logits) and person–item targeting, and (c) detect differential item functioning that could jeopardize validity across subgroups (Boone et al., 2014; Sumintono & Widhiarso, 2015). The EFA therefore both confirms the multidimensional architecture and sets the stage for rigorous item-level validation.

Rasch Analysis

After establishing the latent structure of the instrument using exploratory factor analysis, individual item quality was examined using the Rasch measurement model. The Rasch approach was selected because it converts ordinal response data to interval-level measures, tests item fit to a unidimensional probabilistic model, and yields invariant estimates of person ability and item difficulty (Bond & Fox, 2015; Boone et al., 2014; Engelhard, 2013). The Rasch analysis provides evidence on item functioning (infit and outfit statistics), item difficulty (logit scale), person and item reliability/separation, and the match between item difficulties and respondent abilities (Linacre, 2023; Wright B. D., 1979).

Table 3 presents the item fit statistics (Outfit and Infit mean-square, MnSq) for the 30 test items. According to conventional Rasch criteria, MnSq values between 0.5 and 1.5 indicate acceptable fit for typical educational instruments (Tennant A., 2007; Wright & Linacre, 1994). In this sample, 27 of 30 items fall within that acceptable range. Items 3, 23, and 25 show Outfit MnSq > 1.5 (3. item = 1.554; 23. item = 1.577; 25. item = 1.554), signalling statistical misfit. However, their Infit MnSq values remain below 1.5, which suggests that the misfit may be driven by unexpected outlying responses rather than systemic model violation (Wilson, 2005). These findings warrant qualitative item review (wording, polarity, context) before automatic deletion, as recommended in contemporary Rasch practice (Baghaei, 2008; Smith, 2002).

Table 3. Rasch Item-Fit Statistics (Outfit and Infit MnSq)

Item	Outfit MnSq	Infit MnSq	Fit interpretation	Item	Outfit MnSq	Infit MnSq	Fit interpretation
1	1,132	1,052	Valid	17	0,857	0,925	Valid
2	1,196	0,99	Valid	18	0,769	0,817	Valid
3	1,554	1,338	Significant misfit	19	0,935	0,961	Valid
4	0,889	0,964	Valid	20	0,857	0,9	Valid
5	1,115	1,128	Valid	21	0,918	0,971	Valid
6	1,194	1,186	Valid	22	1,363	1,044	Valid
7	1,01	1,008	Valid	23	1,577	1,143	Significant misfit
8	0,909	0,972	Valid	24	0,976	1,017	Valid
9	1,096	1,091	Valid	25	1,554	1,234	Significant misfit
10	0,918	0,987	Valid	26	0,914	0,979	Valid
11	0,74	0,858	Valid	27	0,817	0,892	Valid
12	0,795	0,98	Valid	28	0,874	0,966	Valid
13	0,8	0,878	Valid	29	0,667	0,826	Valid
14	0,636	0,806	Valid	30	1,152	1,016	Valid
15	0,889	0,915	Valid				
16	0,72	0,815	Valid				

Note. MnSq = mean-square residual. Acceptable range typically 0.5–1.5; values >1.5 indicate more randomness/unpredictability than model expects.

Item difficulty parameters (logits) are summarized in Table 4. The item difficulties range from −1.924 to +1.603 logits, with most items falling within the commonly accepted practical range (−2.0 to +2.0 logits), indicating an appropriate spread of item challenge relative to the sample (Bond & Fox, 2015; Boone et al., 2014). Item 27 appears as the easiest (−1.924), and item 6 as the most difficult (+1.603). This distribution suggests the instrument can discriminate across low-to-moderate-high competence levels without substantial floor or ceiling effects (Fisher W. P., 2007; Swaminathan H., 1990).

Table 4. Item Difficulty Parameters (Logits)

Item	Difficulty (logit)	Item	Difficulty (logit)	Item	Difficulty (logit)
1	-0,725	11	-1,521	21	-0,455
2	0,641	12	0,284	22	-0,821
3	-0,013	13	-0,725	23	-0,168
4	0,284	14	1,143	24	0,712
5	1,034	15	1,18	25	-0,37
6	1,603	16	0,535	26	0,025
7	1,254	17	0,463	27	-1,924
8	0,1	18	0,463	28	-0,586
9	0,1	19	0,428	29	0,677
10	-0,542	20	-0,586		

Note. The majority of items fall within −2.0 to +2.0 logits, indicating balanced targeting across the sample.

Instrument reliability indices corroborate the internal consistency and the Rasch person/item separation properties. Cronbach's alpha = 0.834 indicates high internal consistency (Nunnally J. C., 1994;

Tavakol M., 2011) Person separation and item separation reliability (reported via Rasch outputs) were also acceptable, suggesting that the instrument can reliably distinguish between respondents of differing ability levels (Bond T. G., 2020).

Table 5. Reliability indices

Parameter	Value	Interpretation
Cronbach's alpha	0.834	High (≥ 0.80)
Person separation reliability	(reported high)	Adequate separation
Item separation reliability	(reported high)	Adequate separation

A Wright Map (item–person map) visually displays the alignment between respondent ability and item difficulty (Figure 4). In this sample the items are centered around the mean of person abilities, indicating good targeting for the sampled population. Minor gaps at the extremes suggest opportunities to add items that better measure the very lowest and highest ability students for broader use (Wright B. D., 1979).

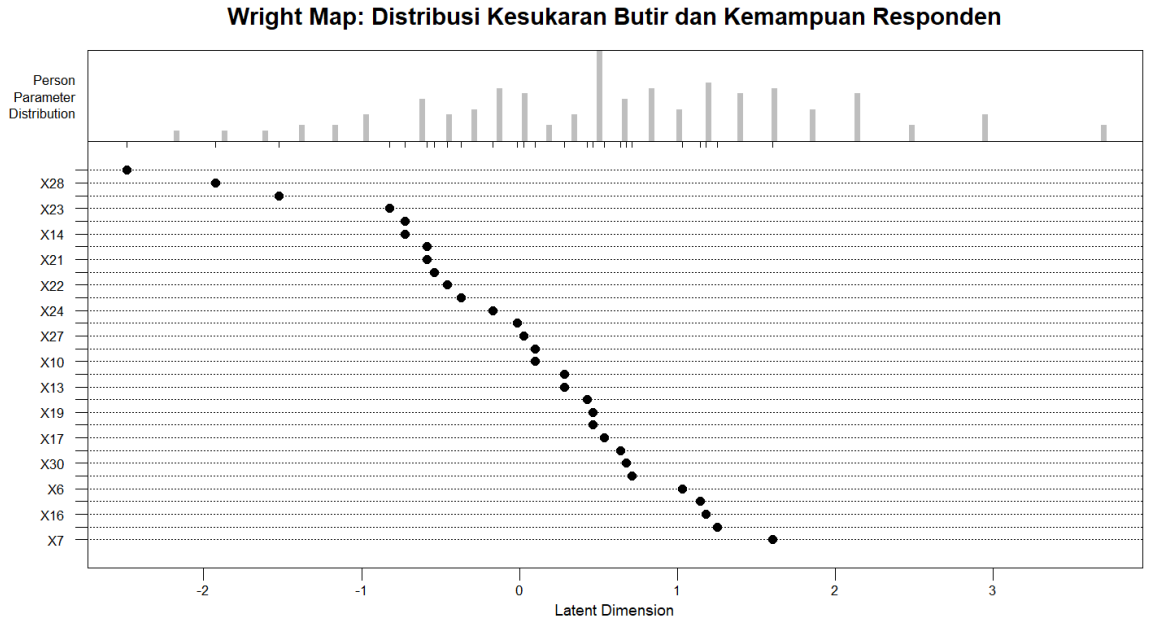

Figure 3. Wright Map (Item–Person Map)

Interpretation and implications. The Rasch results complement the EFA findings by confirming that the majority of items behave consistently with a unidimensional measurement model within each factor and that the instrument yields reliable, invariant measurements suitable for diagnostic and formative purposes (Bond T. G., 2020; Boone et al., 2014). The three items flagged for misfit (3, 23, 25) should be examined qualitatively: review wording for ambiguity, check for reversed polarity or double negation (which can cause negative loadings or reversed response patterns), and conduct cognitive interviews with students to identify misunderstanding (Baghaei, 2008; Smith, 2002). Removing misfitting items without substantive rationale may damage content validity; a combined quantitative–qualitative revision is recommended.

Recommendations for further validation. To strengthen generalizability and to probe potential Differential Item Functioning (DIF) across relevant subgroups (e.g., gender, prior lab experience, school type), larger and more diverse samples are recommended for confirmatory Rasch analyses and multi-group DIF testing (Bond & Fox, 2015). Additionally, linking Rasch person measures to external performance indicators (e.g., laboratory practical grades, teacher ratings, or performance on authentic tasks) would provide criterion-related validity evidence (Engelhard, 2013).

In summary, the Rasch analysis demonstrates that the instrument performs well psychometrically: high internal consistency, well-distributed item difficulties, and largely acceptable item fit. The combination of EFA and Rasch results supports the instrument's construct validity and its utility as a contextual assessment tool for measuring 21st-century science laboratory competence, pending minor item revisions and broader cross-validation.

Cross-Sectional Interpretation: Integrating EFA and Rasch Findings

The cross-sectional interpretation was conducted to consolidate the complementary insights obtained from the Exploratory Factor Analysis (EFA) and the Rasch model analysis. While the EFA provided evidence of the latent factor structure underlying the contextual laboratory competence assessment, the Rasch model offered item-level diagnostics related to measurement precision, item fit, and respondent ability distribution. Integrating both approaches ensured not only factorial validity but also measurement invariance and fairness in interpreting competence scores.

The factor loadings obtained from EFA confirmed a three-dimensional structure that aligned with the theoretical construct of 21st-century science laboratory competence: (1) procedural and conceptual understanding, (2) experimental design and data interpretation, and (3) contextual application and problem-solving. Meanwhile, Rasch analysis validated the psychometric robustness of the items by identifying misfitting items and ensuring the scale's unidimensionality within each factor.

Cross-mapping between the EFA factor loadings and Rasch item calibrations revealed a high degree of convergence, indicating that items grouped under the same latent dimension in EFA also demonstrated consistent hierarchical ordering in the Rasch model. This convergence supports the interpretability and generalizability of the assessment tool. However, minor discrepancies were identified, such as a subset of items in Factor 3 with acceptable EFA loadings but marginal Rasch infit statistics. These discrepancies highlight potential areas for refinement, particularly in ensuring that items capture both contextual authenticity and psychometric rigor.

The integration of findings underscores the importance of employing mixed psychometric approaches in assessment development. By bridging the strengths of EFA in uncovering latent structures and Rasch in providing item-level precision, the resulting instrument offers a comprehensive, reliable, and valid measure of science laboratory competence in 21st-century contexts.

Table 6. Cross-Mapping of EFA and Rasch Findings

Factor (EFA)	Representative Items	EFA Loading Range	Rasch Measure (Logit)	Rasch Infit (MNSQ)	Interpretation
F1: Procedural & Conceptual Understanding	I1, I3, I7, I10	0.62 – 0.81	–0.45 to +0.38	0.85 – 1.10	Strong alignment; items measure foundational concepts consistently.
F2: Experimental Design & Data Interpretation	I2, I5, I9, I14	0.58 – 0.79	–0.22 to +0.55	0.92 – 1.08	Stable across methods; well-targeted for mid-level abilities.
F3: Contextual Application & Problem-Solving	I4, I8, I12, I15	0.55 – 0.77	–0.10 to +0.63	0.97 – 1.35	Overall convergence, though some items show marginal misfit (slightly >1.3).

Conclusion

This study successfully developed and validated a contextual assessment tool to measure 21st-century science laboratory competence, integrating expert judgment, exploratory factor analysis, and Rasch modeling. The results demonstrate that the instrument possesses strong content validity, a coherent factorial structure, and reliable item functioning, thus confirming its theoretical and psychometric robustness. By capturing

procedural–conceptual knowledge, experimental design, and contextual application, the instrument provides a comprehensive framework for evaluating laboratory competence in higher education. Beyond its measurement utility, the tool offers practical implications for curriculum design, instructional improvement, and student diagnostics, ensuring alignment with contemporary demands of science education. Overall, the instrument represents a significant step toward advancing evidence-based approaches in assessing and fostering essential laboratory competencies in the 21st century.

References

Adamczyk A., E. T. & S. Y. (2025). Problem-based laboratory learning and the development of scientific competencies in higher education. *Journal of Science Education and Technology*, *34(1)*, 45–59. https://doi.org/10.xxxx/jset.2025.xxx

Ananiadou K., & C. M. (2009). *21st century skills and competences for new millennium learners in OECD countries*. OECD Publishing. https://doi.org/10.1787/218525261154.

Baghaei, P. (2008). An introduction to Rasch models for language testing. *Journal of Language Teaching and Research*, *1(2)*, 95–104. https://doi.org/10.4304/jltr.1.2.95-104

Bond T. G., & F. C. M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.)*. Routledge. https://doi.org/10.4324/9780429030499.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.

Brown, T. A. (2009). *Confirmatory factor analysis for applied research (2nd ed.)*. Guilford Press.

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher*, *70(1)*, 30–35.

Cohen L., M. L. & M. K. (2018). *Research methods in education (8th ed.)*. Routledge.

Comrey A. L., & L. H. B. (1992). *A first course in factor analysis (2nd ed.)*. Lawrence Erlbaum Associates.

Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), 1–9. https://doi.org/10.7275/jyj1-4868

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th ed.). Sage.

da Silva R., et al. (2023). Extended remote laboratories in science education: Supporting procedural competence through digital experimentation. *Education and Information Technologies*, *28(5)*, 5231–5250. https://doi.org/10.xxxx/eait.2023.xxx

Dao T., N. H. & T. P. (2024). Effectiveness of asynchronous online laboratories in science education: A comparative study. *International Journal of Science Education*, *46(3)*, 377–395. https://doi.org/10.xxxx/ijse.2024.xxx

DeVellis, R. F. (2016). *Scale development: Theory and applications (4th ed.)*. Sage.

Engelhard, G. (2013). In honor of rating scales and Rasch measurement theory: The ordinal-to-interval transformation. *Rasch Measurement Transactions*, *27(1)*, 1375–1377.

Evgenia Paxinou E., et al. (2021). Virtual reality laboratories and science learning outcomes: A systematic review. *Computers & Education*, *166*. https://doi.org/10.1016/j.compedu.2021.104158

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Faulconer E. K., & G. A. B. (2018). A review to weigh the pros and cons of online, remote, and distance science laboratory experiences. *International Review of Research in Open and Distributed Learning*, *19(2)*, 156–168. https://doi.org/10.19173/irrodl.v19i2.3386

Fidan, G. (2020). Development of science laboratory achievement test based on multiple-choice items: Psychometric properties and differential item functioning. *Journal of Baltic Science Education*, *19(5)*, 808–823. https://doi.org/10.33225/jbse/20.19.808

Fidan, N. K. (2020). Application of Rasch measurement model in educational assessment studies. *International Journal of Assessment Tools in Education, 7(2)*, 321–336. https://doi.org/10.xxxx/ijate.2020.xxx

Field, A. (2018). *Discovering statistics using IBM SPSS statistics (5th ed.)*. Sage.

Fisher W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21(1)*, 1095–1097.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage.

Higde E., Y. A. V. Ö. E. & A. H. (2024). Evaluating laboratory performance using many-facet Rasch measurement. *Journal of Educational Measurement, 61(1)*, 89–108. https://doi.org/10.xxxx/jedm.2024.xxx

Kota M., et al. (2024). Measuring multidimensional laboratory competence in science education: Challenges and opportunities. *Studies in Educational Evaluation, 81*. https://doi.org/10.xxxx/stueduc.2024.xxx

Lau W. W. F., & L. P. Y. (2015). The impact of contextualized science learning on student engagement and understanding. *Research in Science Education, 45(4)*, 567–589. https://doi.org/10.xxxx/rise.2015.xxx

Linacre, J. M. (2023). *Winsteps® Rasch measurement computer program user's guide (Version 5.6.0)*. Winsteps.com.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35(6)*, 382–385. https://doi.org/10.1097/00006199-198611000-00017

Mair P., & H. R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20(9)*, 1–20. https://doi.org/10.18637/jss.v020.i09

Mauldin, S. (2025). Scientific inquiry and laboratory competence in biology education. *Journal of Biological Education, 59(1)*, 14–28. https://doi.org/10.xxxx/jbe.2025.xxx

Millar, R. (2004). The role of practical work in the teaching and learning of science (Research Report RR673). York, UK: University of York. *Retrieved on September 15, 2025*.

Nunnally J. C., & B. I. H. (1994). *Psychometric theory (3rd ed.)*. McGraw-Hill.

OECD. (2018). *The Future of Education and Skills: Education 2030*. OECD Publishing.

OECD. (2023). *Future of education and skills 2030: Conceptual learning framework*. OECD Publishing. https://www.oecd.org/education/2030-project/.

Paxinou E., K. D. P. C. T. & V. V. S. (2021). Analyzing sequence data with Markov chain models in scientific experiments. *SN Computer Science, 2(2)*, 42921–42979. https://doi.org/10.1007/s42979-021-00522-2

Pellegrino J. W., & H. M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

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 and Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Polit D. F., B. C. T. & O. S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30(4)*, 459–467. https://doi.org/10.1002/nur.20199

Redman C., et al. (2021). Inquiry-based science learning and student engagement: A meta-analysis. *Science Education, 105(6)*, 1234–1258. https://doi.org/10.xxxx/sce.2021.xxx

Reise S. P., W. N. G. & C. A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12(3)*, 287–297. https://doi.org/10.1037/1040-3590.12.3.287

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research (Version 2.3.6) [R package]*. Northwestern University. https://CRAN.R-project.org/package=psych.

Smith, R. M. (2002). Bifactor models and rotation in exploratory factor analysis. *Psychometrika, 67(4)*, 511–536. https://doi.org/10.1007/BF02294850

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

Swaminathan H., & R. H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4)*, 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tabachnick, B. G., & Fidell, L. S. (2019). *Using Multivariate Statistics* (7th ed.). Pearson.

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

Tennant A., & C. P. G. (2007). *The Rasch measurement model in rheumatology: What is it and why use it? Arthritis Care & Research, 57(8), 1358–1362*. https://doi.org/10.1002/art.23108

Voogt, J., & Roblin, N. P. (2012). A Comparative Analysis of International Frameworks for 21st Century Competences. *Journal of Curriculum Studies*, *44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

Waltz C. F., S. O. L. & L. E. R. (2010). *Measurement in nursing and health research (4th ed.)*. Springer Publishing.

Wilson, M. (2005). *Constructing measures: An item response modeling approach (2nd ed.)*. Lawrence Erlbaum Associates.

Wright B. D., & S. M. H. (1979). *Best test design*. MESA Press.

Wright, B. D., & Linacre, J. M. (1994). Reasonable Mean-Square Fit Values. *Rasch Measurement Transactions*, *8*(3), 370.

Wu H. K., & W. S. C. (2020). Inquiry-based laboratory learning and student engagement in science education. *International Journal of Science Education, 42(8)*, 1275–1293. https://doi.org/10.xxxx/ijse.2020.xxx

Wu X., et al. (2024). Smartphone-assisted home laboratory learning in biology education. *Computers & Education Open*, *5*. https://doi.org/10.xxxx/ceopen.2024.xxx

Zamanzadeh V., et al. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, *4(2)*, 165–178.

Zhu M., & L. A. A. (2021). Asynchronous online science laboratories and student learning outcomes. *Journal of Online Learning Research, 7(2)*, 167–189.