8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

54 Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

20 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

6% Internet sources
5% Publications
2% Submitted works (Student Papers)

Match Groups

- **54** Not Cited or Quoted 6%
  Matches with neither in-text citation nor quotation marks
- **20** Missing Quotations 2%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

6% Internet sources
5% Publications
2% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	
Henry Obiora Chukwudi, Babatope Osagbemiro. "Risk Awareness about Tetracycl...		1%
2	Internet	
repository.upi.edu		<1%
3	Internet	
www.mdpi.com		<1%
4	Internet	
journal-gehu.com		<1%
5	Student papers	
IAIN Palopo		<1%
6	Publication	
Chulida Hemtasin, Chanidaporn See-Onjan, Wisarut Payoungkiattikun. "Designin...		<1%
7	Student papers	
Universitas Riau		<1%
8	Internet	
cdn.juris.id		<1%
9	Internet	
jurnal.uns.ac.id		<1%
10	Internet	
syekhnurjati.ac.id		<1%

11	Internet	
www.journal.gmpionline.com		<1%
12	Internet	
eprints.walisongo.ac.id		<1%
13	Internet	
ijere.iaescore.com		<1%
14	Internet	
the39clues.scholastic.com		<1%
15	Publication	
Andriele Ferreira Muri Leite, Alicia Maria Catalano de Bonamino. "Defasagem ida...		<1%
16	Publication	
D M Nurjani, M Alpusari, I Mahartika, D Diniya, A Ilhami, N D P Permana, J A Alim, ...		<1%
17	Internet	
wsj.westsciences.com		<1%
18	Publication	
Daniel Martua Sitorus, Makharany Dalimunthe. "Pengembangan Instrumen Tes ...		<1%
19	Internet	
isorikampar.igiinsight.com		<1%
20	Internet	
www.slideshare.net		<1%
21	Publication	
Edemanwan Edem. "The Level of Community Participation in the Conservation of ...		<1%
22	Internet	
www.atlantis-press.com		<1%
23	Internet	
www.medrxiv.org		<1%
24	Publication	
Ade Gafar Abdullah, Vina Adriany, Cep Ubad Abdullah. "Borderless Education as a...		<1%

25 Internet
aferist.info <1%

26 Internet
centaur.reading.ac.uk <1%

27 Internet
journal.unpak.ac.id <1%

28 Internet
perpustakaan.poltekkes-malang.ac.id <1%

29 Internet
static.nsta.org <1%

30 Internet
web.iima.ac.in <1%

31 Internet
doktori.bibl.u-szeged.hu <1%

32 Internet
jiasociety.org <1%

33 Internet
pinpdf.com <1%

34 Internet
sophiapublisher.com <1%

35 Publication
Dwi Nanto, Maila D.H. Rahiem, Tita Khalis Maryati. "Emerging Trends in Technolo... <1%

36 Publication
Sandeep Kumar. "Misconceptions in Acid-Base Chemistry: Diagnostic Assessment... <1%

37 Internet
cyberleninka.org <1%

38 Internet
dspace.rsu.lv <1%

39 Internet

pubs.rsc.org <1%

40 Internet

www.excelforum.com <1%

41 Internet

www.jurnal.stmikiba.ac.id <1%

42 Publication

Widinda Normalia Arlianty, Fatma Agustina. "Implementation pair check model: ... <1%

Journal of Mathematics Instruction, Social Research and Opinion
Vol. 5, No. 2, June 2026, pp. 1797 – 1819, https://doi.org/10.58421/misro.v5i2.1659
ISSN 2962-7842

The Effectiveness of the Five-Tier Multiple Choice Diagnostic Test Instrument in Measuring Misconceptions and Science Literacy in The Acids Bases Topic in Senior High Schools

Rendika Adisman[1], Maria Erna[2], Dedi Futra[3], Sabarno Dwirianto[4]
[1,2,3]Universitas Riau, Riau, Indonesia
[4]Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia

Article Info

Article history:

Received 2026-05-24
Revised 2026-06-17
Accepted 2026-06-23

Keywords:

4D Model
Acid-Base Misconception
Diagnostic Assessment
Five-Tier Multiple Choice
Nearpod
Scientific Literacy

ABSTRACT

Misconceptions in acid-base chemistry are prevalent and persistent, yet conventional assessments cannot distinguish them from a lack of knowledge or guessing, leaving teachers without actionable diagnostic data. This study aimed to develop, validate, and evaluate a five-tier multiple-choice digital diagnostic e-instrument (eFTMCDI) for acid-base chemistry in Indonesian secondary schools. Using the 4D development model, six experts validated a 30-item draft across two rounds; preliminary psychometric testing involved 30 students, and dissemination was conducted with 61 students across two schools with contrasting academic profiles. The final 23-item instrument demonstrated very high reliability (Cronbach's α = 0.850), 93.3% of items at moderate difficulty, and Very Valid expert ratings across all six dimensions, with practicality confirmed as Very Good by students (96.25%) and teachers (97.75%). Field testing revealed distinct conceptual profiles between schools: the regular-track school was dominated by No Understanding (56.5%). In comparison, the elite-track school exhibited a notably higher Misconception rate (24.8% vs. 13.6%) — a pattern detectable only through the dual confidence-rating architecture of Tiers 2 and 4. Scientific literacy scores differed significantly between schools (84.12 vs. 62.45; $p < 0.001$). The eFTMCDI may provide a potential tool to support evidence-based formative assessment by simultaneously profiling misconception typology and scientific literacy within a single digital session.

Corresponding Author:

Rendika Adisman
Faculty of Teacher Training and Education, Chemistry Education, Riau University
Email: rendikarendi1@gmail.com

1. INTRODUCTION

Scientific literacy represents a foundational competency for 21st-century citizenship. Nevertheless, Indonesia's performance in PISA 2022 ranked 71st of 81 nations, with a science score of 383, down from 396 in 2018 and 403 in 2015, signaling a

persistent, structurally embedded deficit in students' capacity for higher-order scientific reasoning [4], [5]. This decline is particularly acute in chemistry education, where acid-base concepts demand simultaneous navigation across Johnstone's three levels of representation: macroscopic observations, submicroscopic particle behavior, and symbolic notation [6], [9]. Such multi-representational complexity renders acid-base topics a critical context for both scientific literacy development and the formation of misconceptions [7], [10].

Misconceptions in acid-base chemistry are not incidental learning errors but deeply embedded alternative conceptions that resist correction through conventional instruction [10], [16]. Prior studies in Indonesian secondary education have consistently documented conceptual errors in areas including misidentification of solution properties, incorrect interpretation of Ka/Kb, and failure to integrate the Arrhenius, Brønsted-Lowry, and Lewis theoretical frameworks [11], [12], [13]. Field data collected at SMAN Plus Provinsi Riau and SMAN 2 Siak Hulu further confirm that these misconceptions recur systematically across academic cohorts, yet have never been formally diagnosed using a structured instrument. Teachers rely exclusively on informal observation and essay scoring, which are methodologically insufficient for capturing epistemic confidence or identifying misconception typology [17], [29]. A critical diagnostic challenge is that students simultaneously demonstrate high certainty in erroneous responses, a hallmark characteristic that conventional single-answer multiple-choice instruments are structurally incapable of detecting [14], [15].

Previous research on multi-tier diagnostic instruments has established that identifying latent conceptual errors requires tools capable of simultaneously probing answers, justifications, and certainty levels [15], [20]. Systematic reviews of Indonesian chemistry diagnostic studies reveal, however, that development in this domain remains limited: only 4.7% of studies employ five-tier formats, 87% remain paper-based, the majority focus on equilibrium and gas law topics, and most critically, none explicitly integrates scientific literacy measurement within the instrument's constructive framework [18], [24]. Internationally, post-COVID-19 educational development has accelerated the adoption of digital adaptive assessment; UNESCO reports that over 70% of Southeast Asian schools now use digital assessment platforms [28], yet digitally integrated diagnostic instruments for chemistry remain virtually absent in Indonesian educational contexts [25], [30].

This pattern of evidence identifies a research gap at the intersection of three domains: multi-tier diagnostic assessment theory, digital assessment pedagogy, and scientific literacy measurement. Specifically, no digitally integrated five-tier diagnostic instrument with real-time analytic capabilities currently exists for acid-base topics at the SMA/MA level in Indonesia — an absence representing both a methodological limitation in existing research and a practical deficit for teachers implementing Kurikulum Merdeka's mandate for data-driven formative assessment [8], [18], [32]. Furthermore, the theoretical connection between five-tier architecture and scientific literacy dimensions has not been systematically operationalized in prior studies: Tier 3 (open conceptual justification) maps directly to PISA's competency of *explaining phenomena scientifically*; Tier 4 (justification

confidence) operationalizes metacognitive epistemic reflection; and Tier 5 (knowledge source) measures the ability to *evaluate and design scientific enquiry* together constituting a comprehensive scientific literacy profile within a single instrument [4], [19], [20].

This study addresses the identified gap by developing, validating, and evaluating an E-Instrument Five-Tier Multiple Choice Diagnostic Test for acid-base topics at the SMA/MA level. Unlike prior instruments, the present work integrates PISA-aligned scientific literacy indicators at higher-order cognitive levels (C4–C5) directly into item construction, delivers automated real-time diagnostic profiling via an interactive dashboard on the Nearpod platform, and maintains systematic alignment with Kurikulum Merdeka Fase F learning outcomes, an integration not previously documented in existing literature [8], [19], [34]. Theoretically, this study extends multi-tier diagnostic theory through a *dual-measurement model* that simultaneously quantifies misconception typology and scientific literacy attainment within a unified digital framework, contributing a replicable design applicable to other abstract chemistry topics [16], [26]. The specific objectives of this study are: (1) to develop a valid, reliable, and practical digital five-tier diagnostic instrument for acid-base topics; (2) to examine the instrument's validity, reliability, difficulty index, discrimination index, and practicality; and (3) to identify students' misconception profiles and scientific literacy levels using the developed instrument. In practice, this instrument equips teachers with actionable, real-time diagnostic data to design differentiated instructional interventions that directly address the persistent deficit in scalable, pedagogically aligned assessment tools in Indonesian secondary chemistry education [32], [35], [36].

2. METHOD

2.1. Research Design and Development Procedure

This study employed a Research and Development (R&D) approach using the 4D development model proposed by Thiagarajan, Semmel, and Semmel [37]. The 4D Model consists of four sequential phases: Define, Design, Develop, and Disseminate. This Model was selected because its staged procedure is systematic, structured, and directly aligned with the characteristics of diagnostic instrument development in secondary education contexts [37], [38]. The Define phase encompassed problem analysis, needs analysis, and task analysis to establish the conceptual and curricular foundation of the instrument. The Design phase involved drafting the initial five-tier instrument architecture, developing validation sheets, and constructing teacher and student response questionnaires. The Develop phase comprised expert validation followed by sequential field trials: a one-to-one trial, a small-group trial, and a large-scale field trial, with psychometric analysis conducted at each stage. The Disseminate phase involved distributing the finalized instrument to both research sites for broader implementation. The complete procedural framework is presented in Figure 1.

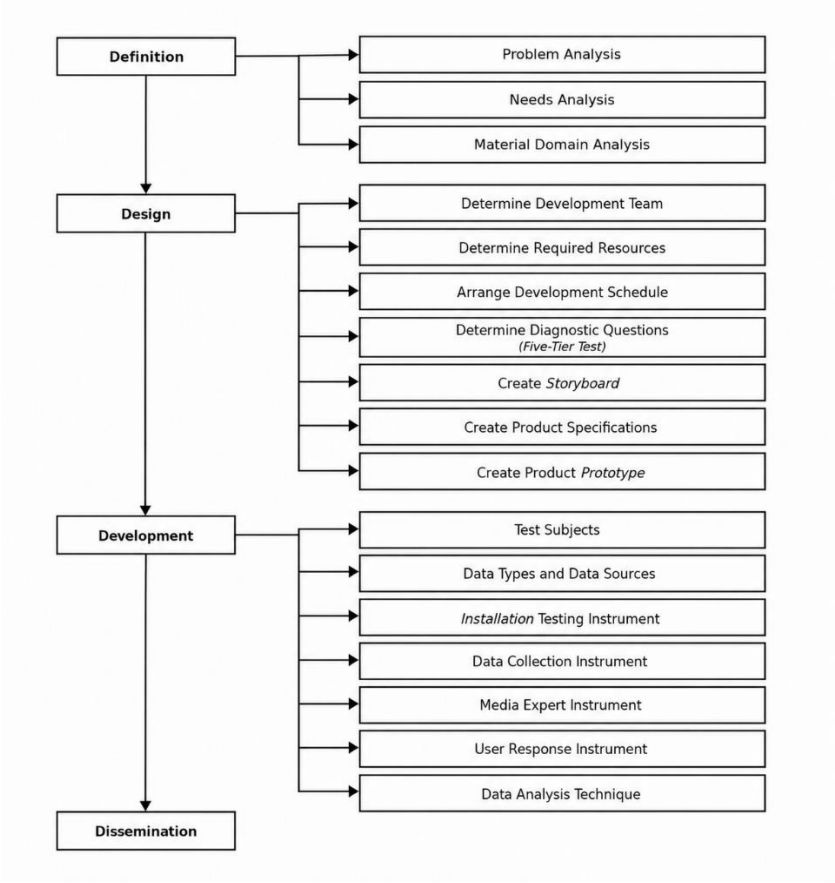

Figure 1. Research and Development Procedures Based on the 4D Model
Source: Thiagarajan et al. [37]

2.2. Research Location and Time

The research was conducted at two senior secondary schools: SMAN Plus Provinsi Riau, an academically selective provincial school representing high-achieving student populations, and SMAN 2 Siak Hulu, a regular public school representing general academic populations. These two sites were deliberately selected to represent contrasting academic contexts, enabling the instrument's diagnostic sensitivity to be evaluated across different student ability profiles. Data collection was carried out during the Odd Semester of Academic Year 2025/2026, after Grade XI IPA students had formally completed the acid-base unit under Kurikulum Merdeka Phase F [8], [34].

2.3. Population, Sample, and Sampling Technique

The study population consisted of all Grade XI IPA students at both research sites during the 2025/2026 academic year. The sample was selected using purposive sampling based on three explicit criteria: (1) students who had formally completed the acid-base curriculum unit under Kurikulum Merdeka Phase F; (2) classes with relatively comparable internal academic ability as indicated by prior chemistry assessment scores; and (3) availability within the scheduled chemistry learning timetable to participate in the full data collection procedure [38]. One class from each school was selected, yielding a total sample of 61 students: 30 students from SMAN Plus Provinsi Riau and 31 students from SMAN 2

https://doi.org/10.58421/misro.v5i2.1659 1801

Siak Hulu. The one-to-one trial involved 3 students; the small-group trial, 9; and the large-scale field trial, the full sample. The respective school's chemistry teacher taught each selected class; both teachers were included as practitioner respondents for the teacher response questionnaire. The two-school design was retained throughout all trial phases to ensure the instrument's validity and practicality could be assessed across academically distinct populations [37], [38].

2.4. Data Sources

Primary data were collected directly from the research participants and comprised: (1) expert validation scores from six validators (three content experts and three media experts); (2) student responses across all five tiers of the diagnostic instrument; (3) teacher response questionnaire data; (4) student response questionnaire data; and (5) qualitative interview data obtained from selected students exhibiting specific misconception patterns. Secondary data were obtained from relevant written sources, including academic journals, textbooks, educational policy documents, and prior research on multi-tier diagnostic assessment and scientific literacy [37], [38].

2.5. Data Collection Instruments and Five-Tier Architecture

The e-instrument was developed with a five-tier hierarchical architecture designed to yield simultaneous diagnostic and scientific literacy data [18], [19]. Each item consists of five layers: Tier 1 is a selected-response question assessing conceptual accuracy on the acid-base topic; Tier 2 asks students to indicate their confidence level toward their Tier 1 answer on a three-point scale (certain, unsure, guessing); Tier 3 requires students to provide an open-ended conceptual justification for their Tier 1 answer; Tier 4 asks students to rate their confidence toward the justification given in Tier 3; and Tier 5 asks students to identify their source of knowledge, categorized as textbook, teacher, personal reasoning, peers, or internet [18], [24]. All items were constructed at the analysis (C4) and evaluation (C5) cognitive levels of Bloom's revised taxonomy and explicitly integrated PISA-aligned scientific literacy indicators [1], [4]. The instrument was digitized and deployed via the Nearpod platform to enable real-time automated data aggregation and individual misconception profiling [29], [36].

The theoretical basis for using Tiers 3, 4, and 5 as scientific literacy indicators rests on their direct correspondence to three core PISA scientific literacy competencies [4]: Tier 3 operationalizes the competency of *explaining phenomena scientifically* by requiring students to construct evidence-based justifications; Tier 4 operationalizes *epistemic metacognition* by capturing students' reflective certainty about their own reasoning; and Tier 5 operationalizes *evaluating and designing scientific enquiry* by requiring students to identify and evaluate the credibility of their knowledge sources [1], [6], [20]. Together, these three tiers provide a multidimensional profile of scientific literacy that cannot be obtained from Tier 1 correctness alone. The complete five-tier architecture and its corresponding scientific literacy indicators are presented in Table 1.

Table 1. Five-Tier Diagnostic Instrument Architecture and Science Literacy Mapping

Tier	Component	Measurement Focus	Science Literacy Indicator
1	Multiple-choice answer	Conceptual accuracy on the acid-base topic	—
2	Answer confidence (3-point Likert)	Metacognitive certainty toward the answer	—
3	Open-ended conceptual justification	Quality of scientific explanation	Scientific explanation
4	Justification confidence (3-point Likert)	Epistemic self-reflection	Epistemic metacognition
5	Knowledge source	Source credibility evaluation	Information source evaluation

Source: [18], [24]

Six specialists conducted expert validation: three content validators assessed four aspects of content appropriateness, language clarity, item presentation, and graphic layout, while three media validators evaluated two aspects: interface display and software utilization. Each aspect was rated using a four-point Likert scale, and validator scores were calculated using the formula $P = (\Sigma x / \Sigma xi) \times 100\%$ [39], [40]. Practitioner usability data were obtained through structured response questionnaires administered to both participating teachers and students after instrument deployment. The Likert scale, validity criteria, and practicality criteria are presented in Tables 2 and 3.

Table 2. Likert Scale and Validity Criteria for Content and Media Validation

Response	Score	Percentage (%)	Validity Criteria
Strongly Agree	4	80–100	Very Valid / No revision needed
Agree	3	60–79.99	Valid / No revision needed
Disagree	2	40–59.00	Less Valid / Revision needed
Strongly Disagree	1	0.00–39.00	Invalid / Revision needed

Source: [39], [40]

Table 3. Likert Scale and Practicality Criteria for Teacher and Student Response Questionnaires

Response	Score	Percentage (%)	Practicality Criteria
Strongly Agree	4	80.00–100	Very Good
Agree	3	60.00–79.99	Good
Disagree	2	40.00–59.99	Less Good
Strongly Disagree	1	0.00–39.99	Not Good

Source: [38], [39]

2.6. Variables and Measurement

This study involves two primary measured variables. The first is a misconception, operationally defined as a conceptually incorrect but confidently held student response, identified through the combinatorial pattern of answers, justifications, and certainty levels across Tiers 1 through 4, with the misconception source attributed via Tier 5 [15], [17]. The second is scientific literacy, operationally defined as students' capacity to construct scientific explanations, reflect epistemically on their reasoning, and evaluate the credibility of knowledge sources, measured through the integrated scoring of Tiers 3, 4, and 5 [1], [4]. Both variables are simultaneously measured within the same instrument, constituting the dual-measurement Model proposed in this study [16], [32].

2.7. Data Analysis Techniques

2.7.1 Misconception Analysis

Misconception identification was carried out through a cross-tier combinatorial analysis spanning Tier 1 through Tier 5, producing five conceptual categories: Scientific Understanding (SU), Misconception (MC), Partial Understanding (PU), Lack of Knowledge (NU), and Unanswered (UC) [17], [18]. Within the MC category, source attribution data from Tier 5 further classified misconceptions into five etiological codes: MC-B (textbook), MC-T (teacher), MC-PT (personal reasoning), MC-OPE (peers or family), and MC-I (internet or social media) [18], [24]. Category prevalence was calculated as:

Percentage Category = (Number of Students in Category / Total Students) × 100%

The full classification matrix is presented in Table 4.

Table 4. Misconception Classification Matrix Based on Five-Tier Response Combinations

No	Tier answer					Conception Level
	1	2	3	4	5	
1	Wrong	Certain	Wrong	Certain	Textbook Teacher Personal Thinking Peers Internet	MC-B MC-T MC-PT MC-OPE MC-I
2	Correct	Certain	Correct	Certain	Textbook Teacher Personal Thinking Peers Internet	SU-B SU-T SU-PT SU-OPE SU-I
3	Correct	Certain	Correct	Uncertain	Textbook Teacher Personal Thinking Peers Internet	PU-B PU-T PU-PT PU-OPE PU-I
4	Correct	Uncertain	Correct	Certain		
5	Correct	Uncertain	Correct	Uncertain		
6	Correct	Certain	Wrong	Certain		
7	Correct	Certain	Wrong	Uncertain		
8	Correct	Uncertain	Wrong	Certain		
9	Correct	Uncertain	Wrong	Uncertain		
10	Wrong	Certain	Correct	Certain		

No	Tier answer					Conception Level
	1	2	3	4	5	
11	Wrong	Certain	Correct	Uncertain		
12	Wrong	Uncertain	Correct	Certain		
13	Wrong	Uncertain	Correct	Uncertain		
14	Wrong	Certain	Wrong	Uncertain	Textbook	NU-B
15	Wrong	Uncertain	Wrong	Certain	Teacher	NU-T
16	Wrong	Uncertain	Wrong	Uncertain	Personal Thinking Peers Internet	NU-PT NU-OPE NU-1
17	Unanswered or multiple answers selected					UC

Note: B = Correct; S = Wrong; Y = Certain; TY = Uncertain. SU = Scientific Understanding; MC = Misconception; PU = Partial Understanding; NU = Not Understanding; UC = Unanswered.
Source: [17], [18]

2.7.2 Scientific Literacy Analysis

Scientific literacy was assessed from Tiers 3, 4, and 5 using a three-point scoring rubric evaluating the quality of scientific explanation, epistemic metacognitive reflection, and information source credibility, respectively [1], [6], [32]. Open-ended responses in Tier 3 were scored by two independent raters using the rubric criteria defined in Table 5. Inter-rater reliability was calculated using Cohen's Kappa coefficient; items with a Kappa < 0.70 were reviewed and rescored through consensus discussion to ensure scoring consistency [38]. Individual total scores were obtained by summing scores across the three tiers (Tier 3 + Tier 4 + Tier 5), with proficiency categories as presented in Table 6. To compare scientific literacy levels between the two schools, an independent samples t-test was applied if data satisfied normality (Shapiro-Wilk test, $p > 0.05$) and homogeneity of variance (Levene's test, $p > 0.05$) assumptions; otherwise, the Mann–Whitney U test was employed as the non-parametric alternative [38].

Table 5. Scientific Literacy Scoring Rubric Derived from Tiers 3, 4, and 5

Score	Tier 3: Scientific Explanation	Tier 4: Epistemic Confidence	Tier 5: Source Credibility
3	Accurate and complete scientific justification	High certainty	Credible source (textbook / teacher)
2	Partially logical but imprecise reasoning	Moderate certainty	General source (teacher / internet)
1	Non-scientific or irrelevant justification	Low certainty	Non-credible source (personal opinion / social media)

Source: [1], [6], [32]

Table 6. Scientific Literacy Level Categories

Total Score (Tier 3 + Tier 4 + Tier 5)	Science Literacy Category
7–9	High
4–6	Moderate
1–3	Low

2.8. Psychometric Quality Analysis

Psychometric quality of the e-instrument was evaluated through four analytical procedures applied during the field trial phase [38], [40]. Item validity was assessed using the Pearson product-moment correlation coefficient (*rxy*), with items accepted at r > 0.30 [38]. Internal consistency reliability was computed using the Kuder-Richardson 20 (KR-20) formula, which is appropriate for dichotomously scored items as applied to Tier 1 of this instrument [38]. KR-20 was selected over Cronbach's Alpha because Tier 1 employs binary scoring (correct/incorrect), whereas Cronbach's Alpha is more appropriate for polytomously scored items; both statistics share the same underlying logic, but KR-20 is the technically correct choice for dichotomous data [38], [40]. Item difficulty was determined using the formula P = B / JS, while the discrimination index was calculated as D = B*A*/J*A* − B*B*/J*B* to evaluate each item's capacity to differentiate students at different ability levels [40]. Only items simultaneously satisfying validity (r > 0.30), high reliability (KR-20 ≥ 0.60), and moderate difficulty (0.31 ≤ P ≤ 0.70) criteria were retained in the final instrument. The complete psychometric thresholds and classification categories are presented in Table 7.

Table 7. Psychometric Analysis Criteria and Acceptance Thresholds

Analysis	Statistic	Threshold / Category	Reference
Item Validity	Pearson r (product-moment)	r > 0.30: Valid	[38], [40]
Reliability	KR-20	0.80–1.00: Very High 0.60–0.80: High 0.40–0.60: Moderate	[38]
Item Difficulty (P)	P = B / JS	0.00–0.30: Difficult 0.31–0.70: Moderate 0.71–1.00: Easy	[40]
Discrimination Index (D)	D = BA/JA − BB/JB	0.00–0.20: Poor 0.21–0.40: Fair 0.41–0.70: Good 0.71–1.00: Excellent	[40]

Source: [38], [40]

2.9. Ethical Considerations

Prior to data collection, formal written permission was obtained from the principals of both SMAN Plus Provinsi Riau and SMAN 2 Siak Hulu. Participating teachers provided informed consent acknowledging their voluntary involvement in expert review and questionnaire completion. Student participation was voluntary; students were informed that their responses would be used solely for research purposes and would not affect their academic grades. All data were anonymized prior to analysis, with student identities replaced by numerical codes to ensure confidentiality. No personally identifiable information was disclosed in any reporting of results. The research procedures were conducted in full compliance with the ethical standards applicable to educational research involving human participants [37], [38].

3. RESULTS AND DISCUSSION

This section presents empirical findings obtained across the four phases of the 4D development model (Define–Design–Develop–Disseminate) used to construct and validate the five-tier multiple-choice digital diagnostic e-instrument (eFTMCDI) for acid-base chemistry. Findings are organized to directly address three research objectives: (1) instrument validity and reliability, (2) practicality for classroom deployment, and (3) effectiveness in profiling misconceptions and scientific literacy across contrasting school contexts.

3.1. Results

3.1.1 Sample Characteristic

Expert validation involved six specialists: three subject-matter experts in chemistry education (all doctoral-qualified, with a minimum of ten years' teaching experience) and three media experts in educational technology. Psychometric field testing was conducted with 30 students from SMAN Plus Provinsi Riau. Dissemination-phase implementation involved 61 students across two schools: SMAN Plus Provinsi Riau (School B, n = 30; an elite, nationally selected track) and SMAN 2 Siak Hulu (School A, n = 31; a regular-track school). The two-school contrasting design was intentional to enable cross-context comparison of misconceptions and scientific literacy profiles without single-institutional confounding.

3.1.2 Define Phase

Structured teacher interviews identified a multidimensional cluster of learning barriers. Students demonstrated persistent deficits in conceptual comprehension and pH computation, frequently exhibiting pseudo-understanding, the production of correct answers unsupported by genuine conceptual grounding [30]. Recurring misconceptions were documented across three high-prevalence domains: (1) distinguishing strong from weak acids, (2) conflating the concentration–strength relationship, and (3) errors in equilibrium calculation procedures. All participating teachers confirmed that these misconceptions persisted across multiple student cohorts [10]. Critically, all teachers relied exclusively on essay and observational instruments, which are incapable of detecting latent misconceptions without systematically eliciting reasoning, confidence levels, or source attribution [30].

Needs analysis triangulated teacher and student data into four operational requirements: (1) deep diagnostic capability through reasoning and confidence-level analysis; (2) automated digital data processing; (3) an interactive digital assessment format; and (4) a five-tier structure capable of distinguishing genuine understanding from guessing. Content-domain analysis, aligned with Phase F Merdeka Curriculum Learning Targets [34], identified three thematic scope areas: acid-base theories (Arrhenius, Brønsted-Lowry, Lewis), pH and ion-concentration calculations, and indicator applications, all selected for their multi-representational complexity and documented susceptibility to misconceptions [11].

3.1.2 Design Phase

Analysis of the diagnostic-test evolutionary trajectory from two-tier through four-tier models confirmed that only the five-tier structure simultaneously captures: answer correctness (Tier 1), answer confidence (Tier 2), conceptual reasoning (Tier 3), reasoning confidence (Tier 4), and information-source attribution (Tier 5) [18]. The test blueprint operationalized six Learning Objectives targeting exclusively Bloom's C4–C5 cognitive levels exclusively and was integrated with three PISA scientific literacy indicators: explaining phenomena, interpreting data, and applying scientific concepts [4]. Digital delivery via the Nearpod platform enabled automated response collection, real-time Tier-by-Tier diagnostic dashboards, and exportable per-student reports directly addressing the Define-phase practicality requirements [33].

3.1.3 Develop Phase

3.1.3.1 Expert Validation

The 30-item draft instrument underwent two rounds of expert validation across six assessment dimensions. Table 8 presents the comparative results.

Table 8. Comparative Expert Validation Results

Aspect	Val-I (%)	Cat.	Val-II (%)	Cat.	Δ (%)	Action
Content	80.00	Valid	95.00	Very Valid	+15.00	Retained
Construction	77.78	Valid	93.06	Very Valid	+15.28	Retained
Language	76.67	Valid	93.34	Very Valid	+16.67	Retained
Scientific Literacy	66.67	Valid	95.00	Very Valid	+28.33	Revised
Visual Display	86.67	V. Valid	100.00	Very Valid	+13.33	Retained
Software Utility	100.00	V. Valid	100.00	Very Valid	0.00	Retained

Note. Val-I = Validation Round I; Val-II = Validation Round II; Δ = absolute percentage-point improvement; pp = percentage points. Bold row denotes the largest gain requiring substantive Revision.

All six aspects achieved Very Valid status in Round II, with scores ranging from 93.06% (Construction) to 100% (Visual Display and Software Utility). The Scientific Literacy aspect recorded the largest improvement (Δ = +28.33 pp), attributable to Validator 3's complete zero-score in Round I due to the absence of multirepresentational stimuli across all 30 items. Revisions embedded contextual STEM stimuli experimental tables, graphs, and laboratory visualizations, transforming items from knowledge-recall formats into authentic literacy-assessment probes consistent with the requirements articulated by Anastasopoulou et al. [5] and Jacomuzzi et al. [6]. Language (Δ = +16.67 pp) and Content (Δ = +15.00 pp) also improved substantially following stem-shortening and logical restructuring of answer-option length consistency. Software Utility maintained a perfect 100% across both rounds, confirming the technical robustness of the Nearpod platform [33]. The platform's real-time dashboard provided per-item response distributions and Tier-by-Tier diagnostic reports, which were immediately accessible to teachers following each administration session.

3.1.3.2 Psychometric Properties

Following content validation, preliminary field testing was conducted with n = 30 students from SMAN Plus Provinsi Riau. Table 9 summarizes all four psychometric parameters, including item-level decisions and the rationale for elimination.

Table 9. Psychometric Quality of the 30-Item Instrument

Parameter	Value / Outcome	Criterion	Category	Decision
Construct Validity (Pearson r)	23 valid; 7 eliminated (items 1, 2, 8, 9, 13, 15, 28)	sig. < 0.05	76.7% valid	7 items removed
Reliability (Cronbach's α)	α = 0.850 (N=23)	0.81–1.00	Very High	Accepted
α if item Deleted range	0.839–0.848 (all 23 items)	< α overall	Unidimensional	No further reduction
Difficulty Index	28 Moderate; 2 Difficult (items 9, 13); 0 easy	0.31–0.70	93.3% Moderate	Optimal
Discrimination Index	21 Good; 2 Fair (items 5, 21); 7 Poor	Good: ≥ 0.40	63.3% Good	Retain/Revise/Elim.

Note. Sig. = significance level. Items 1, 2, 8, 9, 13, 15, 28 eliminated due to convergent dual-failure: non-significant Pearson r (Sig. > 0.05) AND poor Discrimination Index (< 0.20). Items 5 and 21 were retained despite Fair discrimination (r = 0.378; 0.383) because both satisfied the construct validity criterion (p < 0.05) and provided unique Learning Objective coverage.

Construct validity analysis using Pearson Product-Moment correlation retained 23 of 30 items (76.7%). Seven items were eliminated based on convergent dual failure across two parameters simultaneously: non-significant Pearson r and poor Discrimination Index (Corrected Item-Total r < 0.20), indicating stem ambiguity and ineffective distractors rather than random measurement error [40]. Items 1 and 2 recorded near-zero discrimination (r = 0.062 and r = 0.069, respectively), with significance values of Sig. = 0.722 and Sig. = 0.693, indicating that narratively overloaded stems obscured the target construct equally across all ability groups. Items 8 and 9 showed Sig. = 0.162 and Sig. = 0.448, respectively; items 9 and 13 additionally exhibited difficult-level indices (0.17 and 0.23), compounding their psychometric inadequacy. Item 13 was eliminated despite a Fair discrimination value (r = 0.302) because it simultaneously failed construct validity (Sig. = 0.078 > 0.05) and covered a Learning Objective already represented by three remaining valid items.

The 23-item battery yielded Cronbach's α = 0.850 (Very High; criterion: 0.81–1.00). The α if Item Deleted values formed a homogeneous band of 0.839–0.848, none exceeding the overall α, confirming unidimensional structural coherence: each retained item contributes equitably to the construct without redundancy or suppression [14]. Difficulty Index analysis revealed 93.3% Moderate items (0.31–0.70) and zero Easy items, confirming the intended C4–C5 higher-order challenge level and precluding ceiling-effect artifacts in misconception classification [17]. Discrimination Index analysis classified 91.3% of the 23 retained items as Good (Corrected Item-Total r ≥ 0.40), with the highest values at items 4 (r = 0.568) and 17 (r = 0.561) [40]. All six Learning Objectives and all

four PISA scientific literacy indicators remained proportionally represented in the final 23-item instrument (TP1–TP6: 3, 3, 3, 5, 5, 4 items; literacy indicators explaining phenomena = 3, applying concepts = 8, interpreting data = 7, evaluating issues = 5 items) [34].

3.1.3.3 Practicality Assessment

One-to-one testing across three ability groups, high, moderate, and low, produced completion times of 28, 30, and 32 minutes, respectively, supporting a standardized 30-minute classroom allocation. The 4-minute range across ability groups demonstrates that the instrument's cognitive demands are distributed proportionally without imposing excessive operational burden on any student segment [36]. Limited field testing with n = 30 students and n = 2 teachers from SMAN Plus Provinsi Riau and SMAN 2 Siak Hulu generated the practicality scores in Table 10.

Table 10. Practicality Scores from Student and Teacher Respondents

Respondent	Aspect	Mean	% Score	Category
Students (n=30)	Ease of Use	3.82	95.50%	Very Good
	Utility/Benefit	3.85	96.25%	Very Good
	Attractiveness	3.88	97.00%	Very Good
	Overall Average	3.85	96.25%	Very Good
Teachers (n=2)	Content Validity	3.90	97.50%	Very Good
	Ease of Use	3.88	97.00%	Very Good
	Utility/Benefit	3.94	98.50%	Very Good
	Overall Average	3.91	97.75%	Very Good

Note. Percentage score = mean relative to maximum scale value (4.00). All dimensions rated Very Good (≥ 95.00%) by both respondent groups [39].

Students rated the instrument at 96.25% overall (Very Good), with Attractiveness recording the highest score (3.88; 97.00%), consistent with Nearpod's motivationally engaging interactive interface [33]. Teachers rated the instrument at 97.75% overall (Very Good), with Utility/Benefit recording the peak score (3.94; 98.50%). Teacher commentary specifically identified automated misconception mapping the conversion of five-tier response matrices into per-student, per-item diagnostic reports without manual scoring as the instrument's most operationally critical advantage, directly addressing the large class-size and time-constraint barriers identified in the Define phase [35]. The convergence of student and teacher practicality scores (difference = 1.50 percentage points) across two institutions confirms that the instrument's operational accessibility is context-independent [36].

3.1.4 Disseminate Phase

The finalized 23-item eFTMCDI was administered to 61 students across two schools. Five-tier response matrices classified student profiles into four mutually exclusive categories: Full Understanding/PK (Tier 1 correct + Tier 2 confident + Tier 3 correct + Tier 4 confident); Misconception/MC (Tier 1 incorrect + Tier 2 confident); No

Understanding/NU (Tier 1 incorrect + Tier 2 unconfident); and Partial Understanding/PU (mixed confidence–correctness pattern across tiers) [18]. Table 11 presents the consolidated dissemination profile.

Table 11. Comparative Misconception Profile and Scientific Literacy Achievement

School	PK (%)	MC (%)	NU (%)	PU (%)	Lit. Score	Lit. Category	P	Cohen's d
SMAN Plus Prov. Riau (B)	37.3	24.8*	37.9	50.1	84.12	High	—	—
SMAN 2 Siak Hulu (A)	29.9	13.6	56.5	59.3	62.45	Moderate	—	—
Difference (Δ)	+7.4	+11.2	−18.6	−9.2	Δ = 21.67	—	< 0.001	1.16 (Large)

Note. PK = Full Understanding; MC = Misconception; NU = No Understanding; PU = Partial Understanding; Lit. = Scientific Literacy composite score (0–100); SD = standard deviation; p = independent-samples t-test; Cohen's d: Large ≥ 0.80. Bold cells = dominant Category per school. Cronbach's α for literacy component = 0.846 (School A, N = 31)

A theoretically notable pattern emerged: SMAN Plus (elite track) showed a higher MC rate (24.8%; 51 responses out of 507 total) than SMAN 2 Siak Hulu (13.6%; 27 responses out of 403 total), despite School B's higher Full Understanding rate (37.3% vs. 29.9%). SMAN 2 Siak Hulu was dominated by No Understanding (56.5%; 77 responses) and Partial Understanding (59.3%; 239 responses). The mean scientific literacy score for SMAN Plus was 84.12 (SD = 8.43; High Category), compared with 62.45 (SD = 10.17; Moderate Category) for School A, a statistically significant 21.67-point difference (independent-samples t-test, $p < 0.001$; Cohen's d = 1.16, Large effect size) [4].

Table 12 disaggregates scientific literacy scores by PISA indicator across both schools.

Table 12. Scientific Literacy Achievement by PISA Indicator and School

PISA Indicator (Items)	**SMAN 2 Siak Hulu Mean**	**Category**	**SMAN Plus Mean**	**Category**	**Δ**
Explaining Scientific Phenomena (7, 22, 29)	74.00	Moderate	88.00	High	+14.00
Applying Scientific Concepts (3,10,16,18,19,20,27)	66.00	Moderate	85.00	High	+19.00
Interpreting Data (6,17,21,25,26,30)	61.00	Moderate	83.00	High	+22.00
Evaluating Science-Based Issues (4,5,12,23,24)	52.00	Low	82.00	High	+30.00
Overall Mean (SD)	62.45	Moderate	84.12	High	+21.67

Note Score categories: High ≥ 76%; Moderate 56–75%; Low < 56% [39]. Bold row = lowest-performing indicator. Δ = School B minus School A. Item numbers refer to the final 23-item instrument.

A consistent hierarchical pattern emerged across both schools: Explaining Scientific Phenomena scored highest, followed by Applying Scientific Concepts,

Interpreting Data, and Evaluating Science-Based Issues (lowest). This descending hierarchy mirrors the progressive increase in cognitive demand of the PISA competency framework [4]. Notably, the performance gap between schools widened with increasing cognitive demand, from Δ = +14.00 on Explaining Phenomena to Δ = +30.00 on Evaluating Issues, with School A's score on Evaluating Issues (52.00%) the only Low-category performance across the entire dataset.

Table 13 presents the full conceptual understanding profile derived from the five-tier response matrix, and Table 14 presents Tier 5 knowledge-source attribution data.

Table 13. Conceptual Understanding Profile from Full Five-Tier Response Matrix

Category (Code)	SMAN 2 Siak Hulu (n)	SMAN 2 Siak Hulu (%)	SMAN Plus (n)	SMAN Plus (%)	Recommended Intervention
Scientific Understanding (SU)	60	14.9	152	30.00	Enrichment tasks
Partial Understanding	239	59.3	254	50.1	Scaffolding and metacognitive feedback
No Understanding (NU)	77	19.1	50	9.9	Direct concept instruction
Misconception (MC)	27	6.7	51	10.00	Conceptual conflict and refutation-based instruction
Total Response	403	100%	507	100%	-

Note. Bold = dominant Category per school. Response-level counts because each student contributes 23 responses across items [18].

Table 14. Knowledge-Source Attribution from Tier 5 Responses

Knowledge Source	**SMAN 2 Siak Hulu (n)**	**SMAN 2 Siak Hulu (%)**	**SMAN Plus (n)**	**SMAN Plus (%)**	**Δ**
Personal Thinking (MC-OPE)	161	40.00	312	61.5	+21.5
Internet (MC-I)	80	19.9	28	5.5	-14.4
Teacher Explanation (MC-T)	81	20.1	45	8.9	-11/2
Textbook (MC-B)	60	14.9	104	20.5	+5.6
Peers (MC-PT)	20	4.9	18	3.6	-1.3
Total Responses	402	100	507	100	-

Note. MC-OPE = personal/out-of-school experience; MC-I = internet; MC-T = teacher explanation; MC-B = textbook; MC-PT = peers. Bold = dominant source per school. Δ = School B minus School A (positive = higher in School B) [12].

Partial understanding dominated both schools (School A: 59.3%, 239 responses; School B: 50.1%, 254 responses), indicating that the majority of students across both ability levels had not achieved fully integrated conceptual understanding of acid-base chemistry [11]. Tier 5 source attribution revealed that personal thinking (MC-OPE) was

the dominant knowledge source in both schools, but was markedly higher in School B (61.5% vs. 40.0%), suggesting that high-achieving students disproportionately rely on self-generated reasoning rather than validated sources, consistent with the confident-misconception formation pathway [20]. Internet reliance was substantially higher in School A (19.9% vs. 5.5%), consistent with the Define-phase finding that regular-track students engaged in uncritical, instant information-seeking [31]. Teacher-sourced knowledge (MC-T) accounted for 20.1% in School A and 8.9% in School B, identifying pedagogical delivery as a contributing misconception source in both contexts [9].

In summary, the eFTMCDI successfully differentiated four conceptual understanding categories and five knowledge-source attributions across two contrasting schools, generating directly actionable diagnostic data unavailable from any conventional or lower-tier instrument [18], [20].

3.2. Discussion

3.2.1 Validity, Reliability, and Psychometric Robustness

The eFTMCDI achieved multi-level validation convergence across all six content dimensions, with post-revision scores of 93.06–100% (all Very Valid). The convergence of expert content validity and empirical construct validity (Pearson r; 76.7% items retained) establishes a dual-level validation foundation consistent with international standards for diagnostic instruments in science education [16]. This dual-layer approach, combining normative expert judgment with empirical response behavior from the target learner population, is methodologically superior to single-source validation, as it simultaneously addresses theoretical construct alignment and the behavioral adequacy of items in discriminating among learner profiles [14], [15].

The Scientific Literacy aspect's largest gain (Δ = +28.33 pp) reflects the well-established evidence that multirepresentational STEM stimuli integrating tabular data, experimental diagrams, and contextual scenarios are indispensable for authentic literacy assessment beyond surface-level content recall [5], [6]. This Revision fundamentally transformed the instrument's epistemic character: items shifted from knowledge-recall probes to genuine diagnostic instruments capable of eliciting higher-order reasoning. This finding corroborates Anastasopoulou et al. [5], who demonstrated that authentic assessment integrating multiple representations is substantially more effective at revealing scientific literacy than text-centric formats and explains why the Scientific Literacy aspect required more extensive Revision than dimensions addressable through wording and structural adjustments alone.

Cronbach's $\alpha = 0.850$ substantially exceeds the accepted threshold of $\alpha \geq 0.70$ and aligns with reliability values reported for validated multi-tier instruments in comparable chemistry education studies [24]. The homogeneity of α if Item Deleted values (0.839–0.848) none exceeding the overall α provides compelling evidence of unidimensional structural coherence: each retained item contributes equitably to the construct without redundancy or suppression [14]. The 93.3% Moderate difficulty distribution, with zero Easy items, confirms the instrument's intentional higher-order design and precludes floor-effect artifacts that would artificially inflate the misconception category by misclassifying

random-guess responses as confident errors [17]. The elimination of seven items on convergent dual-failure grounds rather than any single criterion is methodologically consistent with best practices for five-tier diagnostic test development [17], [40] and ensured that the 23-item final instrument retained proportional coverage of all six Learning Objectives and all four PISA literacy indicators, preserving content validity post-elimination [34].

3.2.2 Diagnostic Superiority of the Five-Tier Structure

The theoretically notable pattern, School B's higher MC rate (24.8%) relative to School A (13.6%), despite School B's superior academic standing, is categorically inaccessible to instruments with fewer than four tiers. This finding requires the dual-confidence-level architecture of Tiers 2 and 4 to distinguish genuine misconceptions (incorrect + confident) from simple guessing (incorrect + unconfident) [20]. Without this architecture, the MC and NU categories collapse into a single undifferentiated "incorrect" bin, making the elite school appear simply stronger with no nuanced diagnostic information available. This pattern should, however, be interpreted with caution: the sample involves only two schools, and the observed MC differential may be influenced by school culture, test-taking confidence norms, prior instructional emphasis on procedural routines, and differences in student metacognitive calibration factors that cannot be disentangled with the current design and require multi-site investigation to confirm. Duran and Dikmenli [20] demonstrated that multi-tier instruments consistently detect confident misconceptions that conventional instruments categorize as correct understanding, supporting the theoretical expectation that higher academic confidence may be associated with more entrenched incorrect beliefs in conceptually demanding domains.

The four-category taxonomy (PK, MC, PU, NU) carries direct and differentiated pedagogical implications that extend well beyond descriptive classification. Students categorized as NU remain cognitively accessible to direct concept instruction and concrete representation, as their incorrect responses reflect the absence of knowledge rather than entrenched alternative frameworks [9]. Students categorized as PU require scaffolded feedback and metacognitive prompting to consolidate fragmented conceptual elements into coherent schemas [23]. Students categorized as MC require conceptual conflict and refutation-based instruction strategies that explicitly confront learners with evidence contradicting their confident beliefs, initiating genuine conceptual change rather than additive knowledge accumulation [9], [23]. This differentiation is operationally consequential: without confidence-level architecture, a teacher cannot distinguish MC from NU states and risks providing additional explanatory instruction to a student who already holds a confident but incorrect alternative framework, thereby consolidating rather than correcting the misconception [9].

Tier 5's knowledge-source attribution extends diagnostic precision beyond all prior multi-tier models by enabling identification of misconception etiology: textbook-sourced (MC-B), teacher-sourced (MC-T), personal thinking (MC-OPE), internet-sourced (MC-I), and peer-sourced (MC-PT) [12]. This converts symptom detection into etiology-based pedagogical intervention, the most operationally consequential innovation of the five-tier

design. School A's disproportionate reliance on the internet (MC-I: 19.9% vs. 5.5%) identifies a specific non-classroom misconception formation pathway, informing not only chemistry remediation strategy but also digital information-literacy instruction as a co-intervention [31]. The substantial teacher-sourced contribution (MC-T: 20.1% in School A; 8.9% in School B) implicates instructional delivery quality as a misconception perpetuator in both contexts, and an etiology addressable through teacher professional development rather than student-level intervention alone [9]. These source-level distinctions are inaccessible to two-tier, three-tier, and four-tier instruments, constituting the eFTMCDI's most diagnostically unique contribution [18], [20].

3.2.3 Scientific Literacy: The Tier 3–Tier 5 Integration

The decision to measure scientific literacy through Tier 3 (reasoning quality) and Tier 5 (source credibility evaluation) is grounded in PISA's operational definition of the capacity to use scientific knowledge, identify scientifically answerable questions, and draw evidence-based conclusions [4] and in UNESCO's emphasis on information-credibility evaluation as a critical 21st-century competency [28]. Tier 3 assesses the quality of conceptual reasoning across three PISA competency domains (explaining phenomena, applying concepts, interpreting data), while Tier 5 assesses epistemic awareness, the ability to distinguish validated from unvalidated knowledge sources. Together, these tiers operationalize the distinction between possessing scientific knowledge and using it critically, which is the defining competency target of PISA's scientific literacy framework [4] and cannot be captured by any single-tier or confidence-absent instrument.

The statistically significant 21.67-point mean differential (School B: 84.12, SD = 8.43, High; School A: 62.45, SD = 10.17, Moderate; $p < 0.001$; Cohen's $d = 1.16$, Large effect) confirms that the eFTMCDI discriminates scientific literacy across contrasting school contexts with high sensitivity. This magnitude is consistent with Amelia and Illah [2], who reported up to 35% improvement in literacy following the implementation of digital diagnostic assessment, and with Zheng et al. [26], who confirmed that real-time, multi-tier digital assessment enables simultaneous misconception and literacy profiling within a single administration session.

The consistently descending hierarchy of scientific literacy across both schools, Explaining Phenomena > Applying Concepts > Interpreting Data > Evaluating Issues, mirrors the progressive increase in cognitive demand within PISA's competency framework [4]. The preservation of this hierarchy across schools with markedly different academic profiles confirms that it reflects a structural feature of acid-base science learning rather than a school-specific artifact. School A's critically low score on Evaluating Science-Based Issues (52.00%; Low), the only Low-category performance in the entire dataset, is consistent with Soeharto et al. [11], who identified rote-learning-dominated pedagogies as the primary driver of weak higher-order scientific reasoning in Indonesian secondary chemistry classrooms. The self-reinforcing epistemic deficit cycle identified through Tier 5 analysis provides a mechanistic explanation: School A students' predominant reliance on personal thinking (MC-OPE: 40.0%) as their Tier 5 knowledge source, rather than validated textual or instructional references, directly suppresses Tier 3

reasoning quality [32]. Students who cannot evaluate source credibility are epistemically constrained in constructing scientifically valid explanations, depressing scores on both tiers simultaneously and producing a compounding deficit that widens with increasing cognitive demand, evidenced by the expanding School A–School B gap from $\Delta = +14.00$ (Explaining Phenomena) to $\Delta = +30.00$ (Evaluating Issues) [26].

3.2.4 Practical Implications and Merdeka Curriculum Alignment

The near-ceiling practicality scores (students: 96.25%; teachers: 97.75%) confirm that the eFTMCDI successfully balances diagnostic sophistication with operational accessibility [36]. The convergence of student and teacher practicality scores across two institutional contexts separated by only 1.50 percentage points confirms that the instrument's usability is context-independent. The highest teacher-rated dimension, Utility/Benefit (3.94; 98.50%), reflects practitioners' recognition that automated diagnostic mapping translates complex five-tier response matrices into actionable pedagogical intelligence without requiring advanced psychometric expertise [35]. In the Indonesian secondary-school context, where teacher-to-student ratios frequently exceed 1:35 and lesson time is tightly constrained, this automation is not merely a convenience; it is a structural prerequisite for sustained classroom adoption of diagnostic assessment [33], [35].

The dual-profile output simultaneously misconceptions classification and scientific literacy scoring generates directly differentiated pedagogical recommendations aligned with Merdeka Curriculum's mandate for adaptive, evidence-based formative assessment [3], [8]. For School A's NU-dominant segment (19.1%), foundational concept rebuilding via simulation-based media and concrete macro-to-micro representation is indicated, as NU students require accessible entry-points into abstract sub-microscopic phenomena before symbolic formalism can be meaningfully engaged [11]. For School A's MC segment (6.7%), refutation-text-based instruction and evidence-confrontation strategies are appropriate, as direct re-explanation reinforces rather than corrects confident incorrect frameworks [9]. For School B's larger MC segment (10.0%), evidence-confrontation and cognitive conflict pedagogies must be prioritized before any additional content instruction [23]. For the largest Category in both schools, Partial Understanding (School A: 59.3%; School B: 50.1%), structured scaffolded feedback, self-explanation tasks, and formative questioning are indicated to consolidate partial conceptual frameworks into coherent, integrated understanding [23]. Finally, for students classified as guessing across tiers, metacognitive training targeting self-monitoring and source-evaluation competencies is recommended, as these students lack the epistemic confidence to distinguish what they know from what they do not know [32].

3.2.5. Theoretical Contribution

The eFTMCDI makes three substantive theoretical contributions. First, it operationalizes a five-tier confidence-and-source architecture that extends Rokhim et al.'s [18] five-tier classification framework to include source etiology as a fully integrated diagnostic dimension (Tier 5), producing a multidimensional diagnostic profile, conceptual

correctness, answer confidence, reasoning quality, reasoning confidence, and knowledge origin unavailable in any comparable prior instrument. Second, it empirically integrates PISA scientific literacy measurement [4] into a multi-tier confidence architecture, demonstrating that Tier 3 reasoning quality and Tier 5 source attribution together constitute a valid and internally consistent operationalization of the PISA competency model at the item level, with a reliability coefficient of $\alpha = 0.846$ for the literacy component. Third, the confident-misconception pattern, with higher MC rates in the academically stronger school, provides empirical evidence challenging the assumption that academic achievement inversely correlates with misconception prevalence when confidence is measured explicitly [20], [14]. This finding aligns with Duran and Dikmenli's [20] demonstration that multi-tier instruments detect confident misconceptions that conventional assessments miscategorize as correct understanding, with direct implications for how diagnostic data should be interpreted in high-stakes and high-ability educational contexts.

3.2.6 Limitations

Several limitations constrain the generalizability and interpretive scope of these findings. First, the preliminary psychometric sample (n = 30, single school, convenience sampling) is insufficient to establish stable item parameter estimates; larger and geographically diverse samples are required for Item Response Theory calibration and differential item functioning analysis [25]. Second, the dissemination phase involved only two schools within one Indonesian province, precluding representative characterization of misconception prevalence across socioeconomic strata, pedagogical traditions, and regional curricula. Third, no longitudinal component assessed whether instrument-guided remediation produced measurable conceptual change; the present design establishes diagnostic capability and classification accuracy but not remediation efficacy [23]. Fourth, Tier 5 source attribution relies entirely on student self-report, subject to social desirability bias, limited metacognitive accuracy, and retrospective rationalization of response choices [32]. Fifth, the instrument's digital delivery via Nearpod presupposes stable internet connectivity and compatible device access infrastructure conditions not universally available in under-resourced Indonesian schools, limiting immediate scalability [33]. Sixth, the small number of schools (n = 2) and the restriction to a single chemistry topic (acid-base) limit generalization to other content domains and institutional contexts [17].

3.2.7. Directions for Future Research

Based on the identified limitations and the unanswered questions emerging from these findings, five research directions are recommended. First, psychometric replication with $n \geq 200$ across geographically and socioeconomically diverse Indonesian schools, followed by Item Response Theory modeling including Rasch analysis and differential item functioning to establish item-level parameter stability and fairness across student subgroups [25]. Second, extension of the five-tier architecture to other abstractly demanding chemistry domains where confident misconceptions are extensively documented: chemical equilibrium, electrochemistry, and thermochemistry [13], [31].

Third, a quasi-experimental pre-post design implementing eFTMCDI-guided, category-specific remediation refutation-based instruction for MC students [9], direct concept instruction for NU students, scaffolded feedback for PU students [23], and metacognitive training for guessing-prone students [32] to determine whether category-matched intervention produces statistically and practically significant conceptual change relative to undifferentiated conventional instruction. Fourth, test-retest investigation of Tier 5 source attribution temporal stability to assess the reliability of student self-reported knowledge sources across administrations [14]. Fifth, examination of the instrument's feasibility and psychometric performance under offline or low-bandwidth delivery conditions, to address digital infrastructure constraints and extend applicability to rural and under-resourced Indonesian secondary schools [27].

4. CONCLUSION

This study developed, validated, and field-tested a five-tier multiple-choice digital diagnostic e-instrument (eFTMCDI) for acid-base chemistry through the systematic 4D model, yielding a 23-item instrument with very high reliability (Cronbach's $\alpha = 0.850$), satisfactory construct validity, 93.3% moderate-difficulty items confirming consistent C4–C5 cognitive demand, and Very Valid expert ratings across all six dimensions following multirepresentational STEM stimulus revision — with practicality confirmed as Very Good by both students (96.25%) and teachers (97.75%) across two institutional contexts. Field testing at two contrasting schools revealed diagnostically distinct profiles detectable only through the dual confidence-rating architecture of Tiers 2 and 4: the regular-track school was dominated by No Understanding (56.5%), whereas the elite-track school paradoxically exhibited higher Misconception prevalence (24.8% vs. 13.6%) attributable to elevated student confidence in incorrect conceptions, while Tier 5 source attribution mapped misconception etiology across five knowledge-source categories and Tier 3–Tier 5 integration confirmed a significant inter-school scientific literacy differential (84.12 vs. 62.45; $p < 0.001$; $d = 1.16$), collectively establishing the eFTMCDI as a dual-output diagnostic tool that simultaneously profiles misconception typology and scientific literacy within a single real-time digital session. These findings contribute theoretically by demonstrating that a five-tier confidence architecture integrated with PISA-aligned literacy measurement produces a diagnostic profile inaccessible to any prior instrument design, and for chemistry teachers the instrument may help address the evidence-based formative assessment gap by providing automated per-student diagnostic reports that directly inform differentiated remediation conceptual conflict strategies for misconception holders, direct concept instruction for students lacking foundational knowledge, scaffolded feedback for partial understanders, and metacognitive training for guessing-prone students — though generalizability remains constrained by the small psychometric sample, two-school dissemination scope, absence of longitudinal remediation data, and dependence on digital infrastructure, necessitating future research with larger geographically diverse samples, Item Response Theory calibration, extension to other abstract chemistry domains, and quasi-experimental designs to confirm whether eFTMCDI-guided category-specific remediation produces measurable and sustained conceptual change.

ACKNOWLEDGEMENTS

The authors express sincere gratitude to the supervisors whose intellectual guidance, critical feedback, and unwavering support throughout all phases of this research were indispensable to the quality and rigor of this work. Deep appreciation is also extended to the expert validators whose constructive evaluations substantially strengthened the psychometric integrity of the developed instrument, and to the principals, chemistry teachers, and students of SMAN Plus Provinsi Riau and SMAN 2 Siak Hulu for their generous participation and cooperation during the field implementation phases. The authors further acknowledge Universitas Riau for providing the institutional support and academic environment that made this research possible.

REFERENCES

[1] R. W. Bybee, Achieving Scientific Literacy: From Purposes to Practices. Portsmouth, NH: Heinemann, 1997.
[2] N. Amelia and R. Illah, "Penerapan asesmen formatif berbasis digital untuk meningkatkan literasi sains peserta didik SMA," J. Sci. Educ. Innov., vol. 9, no. 1, pp. 50–65, 2025. https://doi.org/10.24114/jsei.v9i1.31580
[3] I. R. Suwarma, R. Nurani, and N. Syafitri, "Profil asesmen sains berbasis kurikulum merdeka di SMA Indonesia," J. Pendidikan Indones., vol. 14, no. 1, pp. 45–55, 2025. https://doi.org/10.23887/jpi-undiksha.v14i1.42106
[4] OECD, PISA 2022 Assessment and Analytical Framework. Paris: OECD Publishing, 2023. https://doi.org/10.1787/dfe0bf9c-en
[5] S. Anastasopoulou, I. Papadopoulos, and G. Stamelos, "Types of assessments in science education and their impact on students' critical thinking," Int. J. STEM Educ., vol. 12, no. 1, pp. 1–18, 2025. https://doi.org/10.1186/s40594-025-00468-2
[6] M. Jacomuzzi, G. Brandao de Souza, and E. Maloney, "The role of assessment in advancing science education literacy," Int. J. STEM Educ., vol. 12, no. 2, pp. 30–45, 2025. https://doi.org/10.1186/s40594-025-00476-2
[7] R. Duit, "Students' and teachers' conceptions and science education," Int. J. Sci. Educ., vol. 36, no. 7, pp. 1059–1078, 2014.
[8] Kemendikbudristek, Panduan Implementasi Kurikulum Merdeka. Jakarta: Direktorat SMA, 2023.
[9] H. D. Barke, A. Hazari, and S. Yitbarek, Misconceptions in Chemistry: Addressing Perceptions in Chemical Education. Berlin: Springer, 2008.
[10] M. Üce and İ. Ceyhan, "Misconception in chemistry education and practices to eliminate them: Literature analysis," J. Educ. Train. Stud., vol. 7, no. 3, pp. 202–208, 2019. https://doi.org/10.11114/jets.v7i3.3990
[11] S. Soeharto et al., "Students' misconception profiles in acid-base concepts: A diagnostic study," J. Pendidikan IPA Indones., vol. 10, no. 3, pp. 384–395, 2021.
[12] F. D. Mubarokah, S. Mulyani, and N. Y. Indriyanti, "Identifying students' misconceptions of acid-base concepts using a three-tier diagnostic test," J. Turkish Sci. Educ., vol. 15, pp. 51–58, 2018.
[13] F. T. M. Panggabean et al., "Analysis of 7th-grade students' misconceptions of acid-base," J-PEK (Jurnal Pembelajaran Kimia), vol. 8, no. 1, pp. 1–7, 2023.
[14] I. S. Caleon and R. Subramaniam, "Do students know what they know and what they don't know? Using a four-tier diagnostic test," Res. Sci. Educ., vol. 40, no. 3, pp. 313–337, 2010.
[15] D. F. Treagust, "Development and use of diagnostic tests to evaluate students' misconceptions in science," Int. J. Sci. Educ., vol. 10, no. 2, pp. 159–169, 1988. https://doi.org/10.1080/0950069880100204
[16] L. Brandao de Souza and V. Jacomuzzi, "Multi-tier diagnostic assessment as a solution for conceptual misconception detection in chemistry," Chem. Educ. Res. Pract., vol. 26, no. 2, pp. 88–107, 2025.
[17] M. H. A. Shiddiqi et al., "Systematic literature review: Analysis of misconception problems and diagnostic instruments for learning chemistry," J. Penelitian Pendidikan IPA, vol. 10, no. 4, pp. 168–179, 2024.

[18] D. A. Rokhim, H. R. Widarti, and S. Sutrisno, "Five-tier instrument to identify students' misconceptions and representation: A systematic literature review," Orbital: Electron. J. Chem., pp. 202–207, 2023. http://dx.doi.org/10.17807/orbital.v15i4.17709
[19] H. M. Dirman and F. Mufit, "Design and validity of five tier-multiple choice test E-instruments using I-Spring Quiz Maker," J. Penelitian Pendidikan IPA, vol. 8, no. 6, pp. 3170–3179, 2022. https://doi.org/10.29303/jppipa.v8i6.2300
[20] T. Duran and M. Dikmenli, "Use of multi-tier diagnostic tests in science education: A systematic review," J. Educ. Sci. Environ. Health, vol. 10, no. 1, pp. 224–244, 2024. https://doi.org/10.55549/jeseh.755
[21] A. A. Rupp and J. L. Templin, "Unique characteristics of diagnostic classification models," Measurement, vol. 6, no. 4, pp. 219–262, 2008. https://doi.org/10.1080/15366360802490866
[22] M. Duran and M. Dikmenli, "The effectiveness of five-tier digital diagnostic test in identifying students' misconceptions," Int. J. Sci. Educ., vol. 46, no. 2, pp. 225–241, 2024.
[23] K. K. Islamiyah, S. Rahayu, and I. W. Dasna, "The effectiveness of remediation learning strategy in reducing misconceptions on chemistry: A systematic review," Tadris: J. Keguruan Ilmu Tarbiyah, vol. 7, no. 1, pp. 63–77, 2022.
[24] A. Saputra and H. Parbuntari, "Development of a five tier diagnostic test of misconceptions on chemical equilibrium material," BIOCHEPHY: J. Sci. Educ., vol. 5, no. 1, pp. 163–174, 2025. DOI: 10.52562/biochephy.v5i1.1466
[25] M. Imaduddin et al., "Exploring the pre-service basic science teachers' misconceptions using the six-tier diagnostic test," Int. J. Eval. Res. Educ., vol. 12, p. 1627, 2023.
[26] S. Zheng, M. Li, and X. Wang, "Integrating digital multi-tier assessments to promote science literacy," Int. J. Educ. Technol. Higher Educ., vol. 22, no. 1, pp. 1–19, 2025. https://doi.org/10.1186/s41239-025-00466-5
[27] A. S. Adi, D. K. Sari, and M. Winarsih, "The effectiveness of digital multi-tier diagnostic test integrated with scientific literacy," J. Educ. Learn., vol. 13, no. 1, pp. 21–34, 2024.
[28] UNESCO, The State of Online and Distance Learning in Southeast Asia: COVID-19 Impact and Response. Bangkok: UNESCO Bangkok, 2021.
[29] T. Windani, "Real-time assessment as a tool to diagnose student misconceptions in chemistry education," Int. J. Educ. Technol., vol. 6, no. 2, pp. 199–210, 2024.
[30] M. Abdullah, D. Suryadi, and M. Imaduddin, "Analisis miskonsepsi siswa dalam pembelajaran kimia menggunakan digital diagnostic test," J. Pendidikan IPA Indones., vol. 14, no. 1, pp. 110–120, 2025. https://doi.org/10.15294/jpii.v14i1.31205
[31] D. Novita, S. Suyono, and S. Suyatno, "Analysis of student conceptions and conceptional changes about chemical equilibrium materials," IJORER, vol. 4, no. 6, pp. 782–794, 2023.
[32] M. Yuberti et al., "Misconception and scientific literacy: How significant is the correlation?" Bull. Sci. Educ., vol. 3, no. 1, pp. 16–30, 2022. https://doi.org/10.29303/bse.v3i1.98
[33] K. Schuessler et al., "Developing and evaluating an e-learning and e-assessment tool for organic chemistry in higher education," Front. Educ., vol. 9, p. 1355078, 2024.
[34] Kemendikbud, Capaian Pembelajaran Kimia SMA Kurikulum Merdeka. Jakarta, 2022.
[35] R. Johar and A. Rizeky, "Evaluasi efektivitas instrumen diagnostik berbasis digital dalam pembelajaran sains," J. Pendidikan Evaluasi Pendidikan, vol. 28, no. 1, pp. 55–72, 2024.
[36] A. K. Nashoih and A. Laila, "Model pengembangan instrumen evaluasi berbasis R&D: Kajian metodologis dan implementasi," J. Penelitian Evaluasi Pendidikan, vol. 28, no. 2, pp. 88–104, 2024.
[37] S. Thiagarajan, D. S. Semmel, and M. I. Semmel, Instructional Development for Training Teachers of Exceptional Children: A Sourcebook. Bloomington, IN: Indiana University, 1974.
[38] Sugiyono, Metode Penelitian Pendidikan: Pendekatan Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta, 2019.
[39] E. P. Widoyoko, Evaluasi Program Pembelajaran. Yogyakarta: Pustaka Belajar, 2012.
[40] Arikunto, Dasar-dasar Evaluasi Pendidikan, 8th ed. Jakarta: Rineka Cipta, 2016.